Predictive Models for Shop Floor Optimization in the Agrofood Industry

Inês Maria de Macedo Pinto Ferreira

Dissertação de Mestrado

Orientador: Pedro Amorim
Orientador Externo: Raquel Oliveira

Mestrado Integrado em Engenharia e Gestão Industrial

28-06-2021

Abstract

Within the food industry and high-volume production, it is important to ensure that the marketed products comply with specified quality standards and requirements, given that small quality breaches can easily tarnish the brand image.

This dissertation comprises an industrial optimization project developed in a flour and vegetable oil manufacturing company. In line with the emerging Industry 4.0 technologies, the project entails a plan for strategic growth and operational optimization, based on the concept of digital transformation.

In the industry of flour and vegetable oils, the concentration of commercial solvent present in the final product is considered a fundamental quality feature subjected to strict legislation and tabulated standards. Fluctuations in this measure pose a serious issue that can result in the infliction of penalties for non-compliance with the law, or profit loss due to wasted product. One of the main challenges faced by the business under study is to achieve stable and consistent solvent concentrations in the final product, that meet the legislated safety and quality specifications.

The main goal of this project is to empower the real-time monitoring and control of the solvent concentration of the product in circulation. This will be achieved with the support of predictive models that assist the definition of the required actions to reduce the exceeded solvent and its variability. A standardized and iterative Data Mining approach will be leveraged for this purpose, supporting the generation of knowledge concerning the different variables and how they influence the production process. This study focuses on the desolventization equipment that contains multiple sensors that capture data required for the development of the analytical models. Furthermore, the equipment is suited to expand and boost the development of other models in the future in a fast and scalable way.

Mindful of the possibilities for future applications, the projected predictive models not only estimate the hexane concentration of the product in circulation in real-time, but also indicate the optimal thresholds for each operational parameter, according to the required range of solvent concentration. This empowers the prospect of future operationalization by automating alerts, bearing recommended actions to be performed throughout the process. As such, an architecture based on SAP technologies was designed to implement the developed solutions. The predictive models were thus implemented in one of the proposed technological solutions, resulting in the creation of four central dashboards that provide critical information required to successfully monitor and calibrate the process in real-time.

In the end, the results corroborated the effectiveness of implementing predictive models to optimize the solvent extraction process, leading to an improvement in the quality and safety indexes and an enhancement on the information management, supportive of data-driven decision making. The implementation of the proposed architecture leverages Industry 4.0 technologies and artificial intelligence to improve process efficiency and centralized management, ultimately boosting the capabilities of the different process stakeholders.

Resumo

No âmbito da indústria alimentar e produção de elevado volume, é importante garantir que os produtos comercializados cumpram os padrões e requisitos de qualidade especificados, uma vez que, pequenas violações na qualidade, podem facilmente prejudicar a imagem de marca.

A presente dissertação engloba um projeto de otimização industrial desenvolvido numa empresa de produção de farinha e óleo vegetal. Enquadrando as tecnologias emergentes da Indústria 4.0, o projeto prevê um plano de crescimento estratégico e otimização operacional, baseado no conceito de transformação digital.

Na indústria de farinha e óleo vegetal, considera-se a concentração de solvente comercial presente no produto final um parâmetro fundamental, com impacto severo na qualidade do produto, estando esta sujeita a uma legislação rígida, com padrões tabelados. As oscilações desta medida de concentração refletem implicações consideráveis, pelo facto de poderem acarretar à prática de sanções relativas ao incumprimento das normas, ou a uma perda de lucro proveniente do desperdício de produto. Um dos maiores desafios enfrentados pela empresa estudada é o de cumprir as especificações legisladas, relativamente às normas de segurança e qualidade, atingindo ao mesmo tempo, níveis estáveis e consistentes de concentração de solvente no produto final.

O principal objetivo deste projeto é possibilitar a monitorização em tempo real da concentração de solvente no produto em circulação. Deste modo, o desenvolvimento de modelos preditivos servirão de suporte na definição das ações necessárias para a redução da variabilidade e excedente de solvente. Uma abordagem padronizada e iterativa de *Data Mining* servirá como base e fonte de conhecimento relativamente à influência das diferentes variáveis no processo de produção. Este estudo centra-se no equipamento de dessolventização, onde estão instalados vários sensores que captam os dados necessários para o desenvolvimento dos modelos analíticos. Trata-se também de um equipamento capaz de alavancar outros modelos no futuro de forma rápida e escalável.

Atendendo às possibilidades de aplicações futuras, os modelos preditivos projetados não só estimam a concentração de solvente do produto em circulação em tempo real, mas também indicam os limites ótimos para cada parâmetro operacional, de acordo com o intervalo de concentração de solvente que se pretende alcançar. Deste modo, os modelos viabilizam uma operacionalização futura, através da automatização de alertas e de recomendações de ações de ajuste do processo. Assim, desenhou-se uma arquitetura baseada em tecnologias SAP para suportar a implementação das soluções desenvolvidas. Em consequência, os modelos preditivos foram embebidos numa das soluções tecnológicas propostas, resultando na geração de quatro painéis centrais que fornecem informação crítica que viabiliza a monitorização e afinação do processo em tempo real.

Em suma, os resultados corroboraram a eficácia da implementação de modelos preditivos na otimização do processo de extração de solvente, revertendo para uma melhoria nos índices de qualidade e segurança, assim como na gestão da informação que apoia a tomada de decisão. Deste modo, a arquitetura proposta incorpora as capacidades da inteligência artificial com outras tecnologias provenientes da Indústria 4.0, que contribuem para a uma evolução da eficiência do processo, exponenciando as capacidades cognitivas dos seus diferentes intervenientes.

Acknowledgments

I would like to thank my supervisor, Professor Pedro Amorim, for all his support, thoughtful comments and crucial recommendations in the course of this project. I am truly grateful for all the positive reinforcement, helpful advice and the provided guidance that strongly contributed towards this project's fulfilment.

I would also like to express my gratitude towards my colleagues in Deloitte, Raquel Oliveira and Gonçalo Lemos, without whom this dissertation would not have been the same. Your insightful feedback provided the motivation I needed to sharpen my thinking and bring my work to a higher level. Thank you for all the invaluable support, every step of the way. Thank you for your example of hard work, determination and enthusiasm. And thank you for enlightening me on the thrilling world of Data Science and technology that I have become so fond of, and aspire to learn about so much more in the future.

Finally, I would like to thank my family for their unconditional support all these years. My parents and siblings, for their wise counsel and sympathetic ear. Pedro, for all the patience and heartful assistance in every roadblock. You have all played a crucial role in my personal development and growth as whole, and for that, I am ever grateful.

Contents

1 Introduction **1**
- 1.1 Project Framework and Motivation . . . 1
- 1.2 Company Overview . . . 2
- 1.3 Agrofood Sector in Portugal . . . 3
- 1.4 Objectives . . . 4
- 1.5 Approaching Methodology . . . 4
- 1.6 Dissertation Structure . . . 5

2 State of the Art **7**
- 2.1 Industry 4.0 . . . 7
  - 2.1.1 Internet of Things . . . 9
  - 2.1.2 The Shift to Digital Supply Chains . . . 9
  - 2.1.3 Challenges of Digital Transformation . . . 10
  - 2.1.4 The Driving Force of Digital Transformation . . . 11
- 2.2 Data Mining and Statistical Analysis . . . 12
  - 2.2.1 Statistical Indicators . . . 13
  - 2.2.2 Data Mining and Feature Selection . . . 13
- 2.3 Smart Manufacturing Use Cases . . . 15

3 Business Understanding **17**
- 3.1 Oilseed Processing . . . 17
- 3.2 DT/DC Equipment . . . 18
- 3.3 Operational Parameters . . . 19
- 3.4 Quality Management . . . 20
  - 3.4.1 Current State Analysis . . . 20
  - 3.4.2 Future State Proposal . . . 21
- 3.5 Business Goals . . . 22

4 Model Development **23**
- 4.1 Data Understanding . . . 23
  - 4.1.1 Data Presentation . . . 23
  - 4.1.2 Exploratory Data Analysis . . . 25
- 4.2 Data Preparation . . . 28
  - 4.2.1 Data Aggregation . . . 28
  - 4.2.2 Outlier Treatment . . . 29
  - 4.2.3 Dataset Construction . . . 29
  - 4.2.4 Correlation Analysis . . . 30
- 4.3 Modelling . . . 30
  - 4.3.1 Selection of the Modelling Technique . . . 30

4.3.2 Test Design 31
4.3.3 Model Construction 32
4.3.4 Model Assessment 33
4.3.5 Feature Importance 35
4.3.6 Operational Model 36
4.4 Evaluation 37

5 Technology Meets Deployment **39**
5.1 Functional Architecture 39
5.2 Reporting 40
5.2.1 Primary Dashboard 41
5.2.2 Laboratory Analyses Dashboard 42
5.2.3 Operational Parameters Dashboard 43
5.2.4 Parameters' Evolution Dashboard 44

6 Conclusion and Future Work **45**
6.1 Main Conclusions 45
6.2 Future Work 47

A Agrofood Sector, Oilseed Process and Business Understanding **55**

B Model Development Documentation **59**

Acronyms and Symbols

AGV	Automatic Guided Vehicle
AI	Artificial Intelligence
CPS	Cyberphysical Systems
DC	Dryer-Cooler
DMC	Digital Manufacturing Cloud
DSN	Digital Supply Network
DT	Desolventizer-Toaster
ERP	Enterprise Resource Planner
IIoT	Industrial Internet of Things
IoT	Internet of Things
IT	Information Technology
KNN	K-Nearest Neighbour
KPI	Key Performance Indicator
ML	Machine Learning
MLOps	Machine Learning Operations
PDP	Physical-to-digital-to-physical
RFID	Radio Frequency Identification
SAC	SAP Analytics Cloud
SSOT	Single Source of Truth

List of Figures

1.1 Phases of the CRISP-DM Reference Model (Kerber et al., 2000) 5

2.1 The Physical-to-Digital-to-Physical Loop, adapted from Mussomeli et al. (2016) 8
2.2 Digital Supply Chain Transformation (Mussomeli et al., 2016) 10

3.1 As-Is Process 21

4.1 Histograms and boxplots for *PT_gasOut_1* and *PT_floor_5* 26
4.2 Time-series analysis of *SC_fan_10* and *PT_floor_12* 27
4.3 Time-series analysis of *TT_floor_9* and corresponding setpoint 27
4.4 Sliding Window dataset example 31
4.5 Bar chart representing feature importance in the random forest 36

5.1 Functional Architecture 39
5.2 Primary Dashboard 41
5.3 Laboratory Analysis Dashboard 42
5.4 Operational Parameters Dashboard 43
5.5 Parameters' Evolution Dashboard 44

6.1 User Journey 49

A.1 Agrofood Industry Exports Evolution 2010-2020. Source: FIPA (2020) 55
A.2 Oilseed Processing, adapted from: Kemper (2020) 56
A.3 DT/DC Schematic Representation. Source: Kemper (2020) 56
A.4 Client's DT/DC Supervisory System, SCADA 57

B.1 Correlation Matrix Before Outlier Treatment 59
B.2 Correlation Matrix of Dataset A 60
B.3 Correlation Matrix of Dataset B 61
B.4 Correlation Matrix of Dataset C 62

List of Tables

2.1 Node impurity formulas 14

4.1 Summary of the DT/DC operating parameters 24
4.2 Summary of the setpoint variables 25
4.3 Summary of the laboratory measurements 25
4.4 Statistical summary of a selected number of variables 26
4.5 Comparison between the different k parameters for the KNN method 29
4.6 Validation metrics applied to each model for each dataset 34
4.7 Minimum and maximum values of selected variables in each group 37

B.1 Variables selected by the stepwise regression forward Algorithm 63
B.2 Variables selected by the stepwise regression backward Algorithm 64
B.3 Summary of correlated variables in each dataset 65
B.4 Summary of correlated variables in the sliding window Dataset 65
B.5 Minimum and maximum values for each variable in each group 66

Chapter 1

Introduction

The present study comprises a curricular dissertation project conducted in a business environment, proposing to enhance shop floor operations for a company in the agrofood sector.

This first chapter aims to introduce the project's framework and its context in the world of technological advances, followed by the business and sector overview, as well as the projected goals, methodology and structure inherent to the study.

1.1 Project Framework and Motivation

Technological breakthroughs have revolutionized the average person's social, personal, and work-life, as the incorporation of emerging technologies continues to grow exponentially. While technology shapes the competitive landscape, companies are compelled to create disruptive businesses and consistently deliver exceptional experiences to earn and maintain loyal customers. Those who refuse to adopt lean, optimized, and connected technologies essentially refuse to focus on efficiency and business agility, which will eventually result in their demise (Burke, 2020).

The urgency to adapt and respond to dramatic adversity is unquestionable, having been stressed with the Covid-19 outbreak. The arrival of a worldwide pandemic forced many organizations to accelerate and enhance their path towards a digital transformation, augmenting their response to consumers' fluctuating demands (Fui-Hoon Nah and Siau, 2020). In a post-Covid world, people's expectations are bound to change dramatically, as they become used to seamless, quick and efficient deliveries, brought by the new digital models (Uzzaman, 2020). Effectively leveraging trending technologies and building IT-centered business models has thus become a key priority for many leaders (Nofal, 2019). However, the integration of the emerging trends requires a full-on transformation of the business mindset; organizations must be prepared to successfully exploit these technologies for their own benefit. As such, managers are expected to merge their corporate and technology strategies, seeking organizational agility, scalability, and stability (Deloitte, 2021).

There are several technological trends entailing the scope of this dissertation that empower organizations to meet consumers' increasingly high standards. To begin with, cybersecurity plays an

imperative role in data-driven organizations, as companies become increasingly reliant on technology. Currently perceived as the "new oil," data has become one of the most crucial and valuable resources worldwide, and the impacts of a security incident are thus greater than ever (Nofal, 2019). On top of that, artificial intelligence (AI) and industrial automation technologies are expected to flourish in the upcoming years. Specialists are thus expecting Machine Learning Operations (MLOps) to automate the development of machine learning (ML) models. The goal is to empower the automation of manual, inefficient workflow and streamline all the steps of model construction, shifting the focus of AI teams away from model building and towards operationalizing (Deloitte, 2021). Organizations will be prone to automate anything that can possibly be automated, using AI, ML, and robotic process automation (Burke, 2020). This concept is referred to as hyper-automation; it enhances employee productivity by automating time-consuming tasks and increasing operations' flexibility and scalability. Finally, organizations are likely to change the way they capture, store and process information. Managing organized, clean data for human consumption will no longer be their sole focus, as cloud data warehouses are being leveraged to store extensive volumes of unstructured data and feed AI and ML tools (Deloitte, 2021). These technologies are trending exponentially due to their ability to collect large amounts of data from multiple sources, allowing users to search, analyze and mine the data in real-time (Kerner, 2019).

The present curricular dissertation project, deployed in a business environment, reviews the operational enhancement of a company in the agrofood industry, whose main activity is the production of flour and vegetable oils. Food quality and safety regulations represent a crucial aspect of the agrofood industry, as this particular segment is susceptible to significant governmental laws and interventions. The manufacture of flour and vegetable oils requires rigorous monitoring of multiple variables, processes and products, as a means to guarantee the legislated quality specifications. Hence, the use of advanced analytics can be leveraged to monitor, control and predict several parameters, such as the commercial solvent concentration of the final product, which is this project's main focus. In fact, due to the processing and calculation capabilities inherent to AI and advanced analytics, human capacity can be extended beyond its natural cognition. Both worlds must coexist and complement one another, as business knowledge is a fundamental tool required to interpret the end results and turn raw information into valuable insights.

In response to the exponential rise of Industry 4.0 technologies, a selection of ML techniques will be explored in this dissertation, aiming to increase the quality indicators of the company subject to study. The client is highly affected by the instability of the solvent extraction process, and the excess of commercial solvent in the extracted flours was identified as a critical, costly problem. A mishandling of this issue can translate into heavy expenses and harmful loss of competitiveness, hence the urgent need to tackle the subject.

1.2 Company Overview

This project was carried out in Deloitte Touche Tohmatsu Limited (DTTL), commonly known as Deloitte. The firm is a multinational, global leader that provides services in audit, tax, consulting,

and financial advisory, covering a wide range of industries. Currently one of the Big Four accounting organizations, Deloitte is the largest professional services network in the world in terms of revenue and number of professionals (Statista, 2021).

Technology and creativity establish the foundation for Deloitte's consulting services, as they focus on improving their clients' connection with the business, providing an integrated, adaptive, end-to-end journey through the process of business transformation. The present dissertation project integrates the Enterprise Applications team in the Business Intelligence segment, one of the firm's many consulting areas. Enterprise Applications can impact multiple aspects of an organization, enabling businesses to tackle their most complex challenges by building an information-based management culture (Deloitte, 2020).

This project was conducted for a client recognized as a great industrial player in the agrofood sector, operating in the oilseed business segment. Their focus lies mainly on the production of flour and vegetable oils.

1.3 Agrofood Sector in Portugal

Over the past few years, the agrofood industry has faced a significant evolution, preserving an important role in the European and Portuguese economy. As a large supporter of national exportation, the industry continues to grow in exports, as shown in Figure A.1 in Appendix A. According to INE (2020), the total exports of goods in Portugal dropped 11,5% in 2020, from January to October, yet for the same period, exports in the agrofood sector demonstrated a steady increase of 6,2%.

The Portuguese agrofood sector is classified among the industries that contribute the most towards national economy, being the second-largest employer in the country and responsible for a turnover of approximately 17 billions (do Campo, 2020). Marked by elevated, increasing levels of competitiveness, the Portuguese agrofood industry is one of the most developed sectors in the country and is now facing exponential growth (SISAB, 2017). Successfully adapting products to consumers' tastes, seeking healthier production processes and introducing innovative features has laid down the industry's path for substantial growth whilst continuously increasing the competitive edge (ENEI, 2014).

Faced with a growing demand for high quality and safe, sustainable products, the agrofood industry continuously strives to keep up with manufacturing regulations, namely the rules on labelling, hygiene and additives. Not only that, but in light of the current pandemic situation, the countless consequences remain unpredictable, as the complexity of upcoming challenges will inevitably escalate. Hence, the need to promote national commitment around this sector is stressed, and the priority should be to invest in research, development and innovation, leveraging external growth to stimulate the industry (do Campo, 2020).

1.4 Objectives

The current dissertation project will contribute towards the implementation of a machine learning model in a business environment, focusing on two main goals:

1. The development of two predictive models, where the first provides a real-time prediction of the final product's quality, and the second depicts the optimum operating parameters to ensure the legislated quality and safety standards are met.

2. The project implementation in SAP Analytics Cloud, a technological solution that embeds the developed analytical models. This aims to increase the overall efficiency and effectiveness of daily tasks carried out across the organization's different hierarchy levels, ranging from top management down to shop floor operators.

The model development will be employed using the programming language Python. Its implementation will enable the real-time monitoring of the solvent concentration, leading to an improvement of the final product's quality and safety indicators.

1.5 Approaching Methodology

This dissertation followed an approach commonly used by data mining experts, called the Cross Industry Standard Process for Data Mining (CRISP-DM). It is designed to address data mining problems in industrial projects, empowering the search for patterns, trends and correlations in a dataset.

This method divides the data mining process into six stages, as shown in Figure 1.1. It is important to note that these phases are not strictly sequential, and agile iterations between each stage are generally required. The output of each phase indicates which tasks or phase should follow; hence the arrows signal the most frequent dependencies between steps. The outer circle symbolizes the cyclical nature of data mining itself since the process is not over once the solution is deployed. The insights gained during the process and after applying the solution can trigger new, often more-focused business questions. As a result, future data mining processes will benefit from the experience of previous projects.

All six phases of the CRISP-DM model were performed in this project and are described as follows:

1. **Business understanding:** The initial phase involves analyzing the current project's objectives and how they relate to the requirements and goals from a business perspective. This knowledge is required to translate it into a data mining problem, which is designed to achieve those objectives.

2. **Data understanding:** This stage begins with the collection of data, followed by a thorough analysis that enhances familiarity with the dataset. This enables the identification of quality issues, unveiling first insights and forming small subsets to shape preliminary hypotheses.

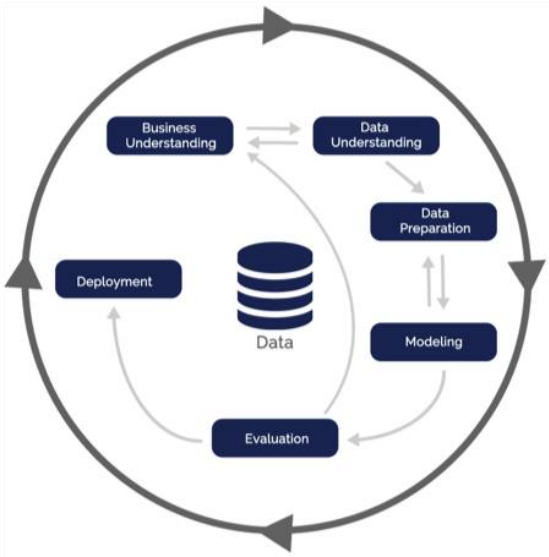

Figure 1.1: Phases of the CRISP-DM Reference Model (Kerber et al., 2000)

3. **Data Preparation:** All the steps required to transform the raw data and construct the final dataset are addressed in this stage. Data preparations tasks are carried out multiple times and include selecting tables, records, and attributes, as well as transforming and cleaning the data to ensure its compatibility with the modelling tool.

4. **Modelling:** In this phase, several modelling techniques are selected, applied and then optimised by calibrating their parameters to ideal values. As there are multiple techniques that can be applied to the same data mining problem, revisiting the preparation phase is often necessary, since some techniques have specific requirements on the format of the data.

5. **Evaluation:** At this point, a model has been built and appears to bear high quality, from a data analysis perspective. However, before proceeding to the final implementation of the model, a thorough evaluation must be carried out. This requires a review of all the steps in its creation and confirming the model achieves the business objectives.

6. **Deployment:** Finally, the knowledge and insights generated by the model are presented to the end-user in a comprehensive and useful manner. The project's solutions are integrated on the shop floor and the client is trained to understand how the model and the obtained results can be used appropriately (Kerber et al., 2000).

1.6 Dissertation Structure

This dissertation is divided into six chapters, starting with the current section that aims to describe the project framework and its goals whilst introducing a brief overview about the company where the project was developed, as well as the client and respective industry.

Chapter 2 exposes a literature review concerning the topics addressed throughout the dissertation, providing the theoretical background that supported the project's development.

Subsequently, Chapter 3 aims to provide an understanding of the business at hand, the oil extraction process, and the equipment used, establishing the alignment with the business goals.

A detailed characterization of the provided data is then presented in Chapter 4, disclosing all the steps required to prepare the data, followed by the ML model development and evaluation.

The deployment phase is then described in Chapter 5, exposing the functional architecture of the employed solution, as well as the implementation techniques and resulting dashboards.

Finally, Chapter 6 provides a reflection of the dissertation's main findings, as well as an exposure of possible future work.

Chapter 2

State of the Art

The following chapter illustrates the imminent evolution of technology, characterizing where it currently stands and where it's potentially headed. The resulting key challenges are described accordingly, followed by a number of relevant use cases for manufacturing processes. In short, the conducted research provided the theoretical foundation required for the development of this project, revealing the importance of Industry 4.0 technologies and their impact on organizational competitiveness.

2.1 Industry 4.0

The need to adapt and respond to change has become an urgent theme for organizations seeking to thrive in today's economy. A statement adapted from Darwin's Origin of Species argues *"it is not the strongest of the species that survives, nor the most intelligent. It is the one that is most adaptable to change"* (North and Varvakis, 2016). This emphasizes the fact that organizations must be flexible and responsive to change in order to survive, as they become faced with an uprise of disruptive technologies, along with the undeniable versatility of consumers' demands (Williams and Olajide, 2020). Technology is a powerful source of competitive advantage, forcing manufacturers to strategically plan their technological investments (Sniderman et al., 2016). Creating new, connected experiences for customers, partners, and workforce, enhances an organization's response to the changing market conditions, laying down the groundwork to earn and maintain loyal customers (SAP, 2020).

The term "Industry 4.0", also addressed as smart manufacturing, refers to a 4th Industrial Revolution, marked by the advanced digitization within organizations (Nicoletti, 2020). The concept is radically reshaping the competitive landscape, having been first introduced in 2010 in Germany, who thus became the most competitive manufacturing country and a global leader in equipment manufacture (Karmakar et al., 2019).

In the course of history, the world faced three major technological shifts that instigated radical change, starting with the 1st Industrial Revolution in the 18th century. This era illustrates one of the greatest turning points in human history, where steam-powered technologies enabled the

development of mechanical production facilities and the massive expansion of several industries (Mohajan, 2019). The 2nd Industrial Revolution followed, starting in the mid-19th century, where the introduction of electricity lead the way for ground-breaking inventions, enabling mass production (Mohajan, 2020). Finally, the 1960s marked the beginning of the 3rd Industrial Revolution, introducing electronic automation. The development of computers and internet connectivity provided the means to not only automate production further, but also to access vital information that considerably improved business management and decision-making capabilities (Rifkin, 2011).

Industrial processes have significantly evolved over time and are now exposed to a digital transformation (Cotteleer and Sniderman, 2017). Industry 4.0 denotes a paradigm shift to a physical-to-digital-to-physical (PDP) connection, as demonstrated in Figure 2.1. The up and coming advanced manufacturing techniques go beyond a simple one-way connection to smart technologies. At the moment, many organizations are already capturing physical information from an object (i.e. product dimensions) to create a digital record (Mussomeli et al., 2016). However, it is the leap from digital *back* to physical that essentially characterizes the concept of Industry 4.0 (Cotteleer and Sniderman, 2017). Through the use of advanced analytics and machine learning, information obtained from a physical object can be processed by several machines, combining real-time data from multiple sources to generate meaningful discoveries. The transformation back to the physical world occurs through algorithms and automation, translating valuable insights from the analyzed data into effective actions that will change the physical environment (Sniderman et al., 2016).

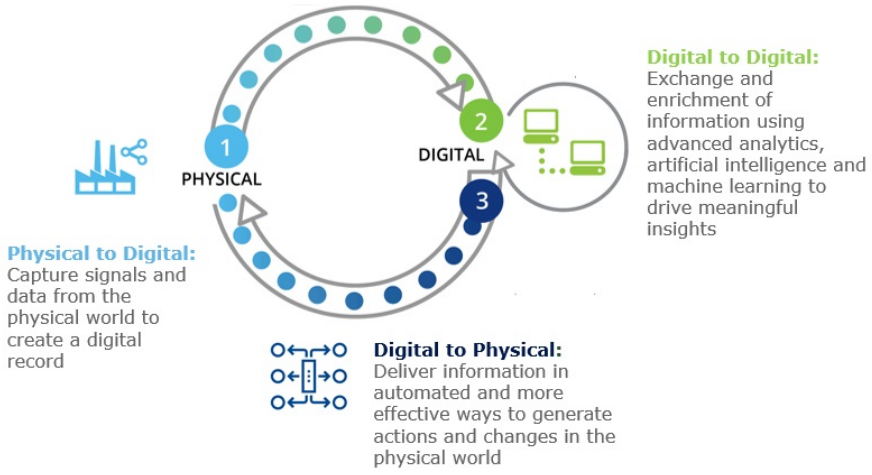

Figure 2.1: The Physical-to-Digital-to-Physical Loop, adapted from Mussomeli et al. (2016)

Even though the basis of Industry 4.0 lies in manufacturing, its implications reach far beyond production processes. Businesses no longer function in a traditional manner; they are compelled to change the way they grasp and manipulate information, not only to achieve operational excellence, but also to continuously improve the customer experience (Cotteleer and Sniderman, 2017). The emerging smart and connected technologies will substantially influence the way products are designed, developed and delivered, as the gathered data enables a deeper understanding of consumer

preferences (Kusiak and Salustri, 2007). This allows companies to efficiently target a specific audience, using customized marketing and selling strategies to change the way consumers interact with their product or service (Cotteleer and Sniderman, 2017)

In the age of Industry 4.0, the physical object is no longer the sole driver of customer experience. Companies can use the fundamentals of the PDP loop to create an interconnected and autonomous digital enterprise, improving consumers' levels of engagement and thereby resulting in increasingly profitable products and services (Cotteleer and Sniderman, 2017). Ultimately, these technologies will revolutionize the rules of production, operations and workforce.

2.1.1 Internet of Things

The Internet of Things (IoT) refers to a network of smart devices, linked together through wireless connectivity, communicating and exchanging data between them (Magomadov, 2020). IoT commercial applications are used to automate and enhance consumers' daily lives, ranging from heart monitors, to autonomous cars, smart-watches and other numerous technological trends. However, while IoT was originally focused on the commercial sector, its undeniable potential lead to an expansion into the enterprise level, resulting in the introduction of the Industrial Internet of Things (IIoT), one of the most prominent elements of Industry 4.0 (Serror et al., 2020).

The IIoT is characterized by Karmakar et al. (2019) as a series of interconnected machines, that through the use of sensors, controllers and other networked devices, provide visibility and insight into a company's operations. It allows businesses to boost productivity, reduce unplanned downtime, deliver high quality and reduce overheads, thus increasing returns on investment (Karmakar et al., 2019). Machine sensors are able to cross-reference their present configuration and environment settings with pre-configured optimal data and thresholds to self-predict, self-compare and become self-aware (Gilchrist, 2016). This way, machines gain the ability to produce self-diagnosis, enabling predictive maintenance to reduce manufacturing disruptions.

Unlike the previous industrial revolutions, triggered by the arrival of a particular technology, Industry 4.0 results from a combination of multiple technologies that together lead to innovative breakthroughs (Parente et al., 2020), such as cyber-physical systems (CPS), IoT, blockchain and cloud computing. In fact, the IoT is the foundation of all Industry 4.0 technologies, being responsible for the connectivity and communication of real-time data between machines (Xu et al., 2018).

2.1.2 The Shift to Digital Supply Chains

According to Mussomeli et al. (2016), the natural growth of Industry 4.0 is causing the transformation of traditional, linear supply chains into dynamic and digital networks, as shown in Figure 2.2. A digital supply chain, also known as a digital supply network (DSN), is described as a flexible, interconnected matrix that enables the flow of data and goods in a nonlinear manner, establishing strong communication between different stages and players of the supply chain (Mussomeli et al., 2016). As the DSN integrates distinct views of the supply network, manufacturers increase

their responsiveness to sudden, radical changes of consumer demands, as they're able to perform last-minute engineering changes, manage inventory digitally and execute data-driven fulfilment decisions (Walsh et al., 2018). As a result, digital supply networks generate important business development opportunities, enabling the reduction of operating costs, improving product quality and increasing visibility and sales effectiveness, which ultimately boosts profitability, while creating strategic advantage (Ozdogru, 2020).

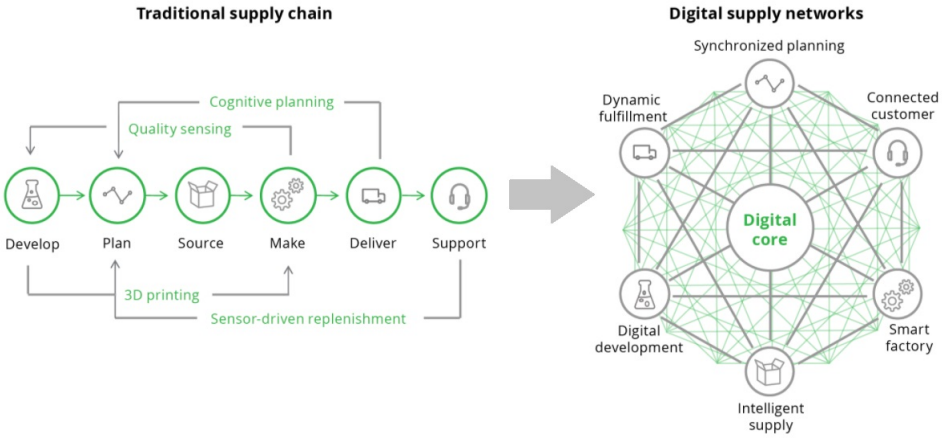

Figure 2.2: Digital Supply Chain Transformation (Mussomeli et al., 2016)

Unlike DSNs, the traditional supply chains, as demonstrated in Figure 2.2, are composed of discrete, sequential operations, starting with the product design and ending in customer support. This linear, ordered approach deprives the company of its chance to effectively react to unforeseen or unexpected events (Choudhury et al., 2021). They are unable to re-route a driver if need be, adapt to sudden weather changes that affect demand forecast, or retrieve accurate inventory data to avoid excessive stock. As such, the shift from a traditional supply chain to an interconnected, open network of supply operations could be the key to overthrow competition in the future (Walsh et al., 2018). Many organizations are already changing their focus away from linear, discrete functions, as they understand the eminent potential of DSNs (Mussomeli et al., 2016). Manufacturers who refuse to evolve and adapt accordingly risk becoming victims of technology disruptions, losing market share and growth opportunities, to those who manage to shift to a responsive, proactive digital supply network (Mahmood, 2019).

2.1.3 Challenges of Digital Transformation

The concept of digital transformation refers to the use of technology to improve overall performance, representing the thrust that reshapes every aspect of modern enterprises (Gilchrist, 2016). Many opportunities arise from smart manufacturing, such as increased productivity, enhanced competitiveness, higher revenue flow and optimized production processes. However, multiple challenges have yet to be addressed to allow Industry 4.0 to thrive to its potential (Xu et al., 2018), including the following issues:

- **Information security** - According to the Ponemon Institute (2018), the biggest challenge in a digitization process is to successfully secure and protect data. Organizations rushing to achieve digital transformation are significantly increasing the risk of a data breach, cyber-attacks and threats to high-value assets. For this reason, cybersecurity must become the primary concern of any organization's transformation plan, focusing on the preservation of confidentiality, integrity and availability of information (Nofal, 2019).

- **Technological infrastructure limitations** - Many manufacturing companies lack the appropriate IT infrastructure to support a digital transformation, presenting gaps of interoperability. In other words, the existing software or computer systems might lack the ability to exchange and make use of information (Gilchrist, 2016). This forces organizations to either reshape their current installations, or invest in brand new infrastructure, implicating substantial costs regardless of the company's size (Sniderman et al., 2016).

- **Data ownership and control** - With an increasing number of stakeholders connected to the value chain, it becomes unclear who owns or controls the exchanged data (Asbroeck et al., 2019). From suppliers and vendors, all the way to the retailers and customers, the interest in the shared data is substantially increasing, as it is considered a strategic asset and a powerful source of value and innovation (ATKearney, 2018). As a result, managers should carefully oversee the contractual agreements of data ownership between the various actors.

- **Shortage of big data skills** - The lack of analytics and data science talent remains a strong barrier for enterprises in the 21st century (Nwokeji et al., 2019). While the number of qualified engineers with big data skills is gradually increasing, it remains insufficient to meet the growing demand of organizations worldwide (Mussomeli et al., 2016). To overcome this shortage of analytical talent, organizations should support and invest in employees' training, aiming towards a digitally sophisticated workforce (Nwokeji et al., 2019).

- **Resistance to change** - Increased resilience to digital change can be observed from top-level management down to shop floor operators. Successful companies are usually the most susceptible to such resistance, as they distrust the need for change when the business runs smoothly. It is thus important for leaders to consider the human side of a digital transformation and focus on tackling these barriers, overcoming traditional mindsets, and implementing small, gradual changes (Scholkmann, 2021).

2.1.4 The Driving Force of Digital Transformation

Organizations are compelled to integrate new technologies in their DSN, such as IoT, AI and ML, to successfully establish a turbulent competitive edge (Ozdogru, 2020). In fact, machine learning has revealed itself to be a prominent research field of AI, expected to drive and discharge growth in the industry, using computer algorithms to uncover patterns in data and accurately predict future events (Anand et al., 2020). New and insightful knowledge generated from ML is revolutionizing

supply chain management, as it enhances operational efficiency by improving demand forecast accuracy, decreasing freight costs and reducing inventory expenditure (Columbus, 2018).

Integrated with the Internet of Things, ML allows a machine to learn by itself due to the rise of big data (Anand et al., 2020), which refers to the huge volumes of structured and unstructured information that organizations currently process and analyze (Nofal, 2019). The increasing availability of data both in size and quality has enabled enterprises to incorporate machine learning and data mining techniques, empowering the extraction of rules from the large quantities of data to support effective decision-making (Dogan and Birant, 2021).

The major approaches of machine learning models can be categorized into supervised learning and unsupervised learning (Anand et al., 2020). Supervised learning deals with labelled data, using regression and classification techniques, where regression is used to predict continuous or ordered values, such as the price of a car, and classification predicts discrete, categorical, or pre-defined values (i.e. small, medium, large) (Dogan and Birant, 2021). On the other hand, unsupervised learning is used to identify regularities and dependencies in unlabelled data, typically using clustering techniques to aggregate objects based on their similarities (Dogan and Birant, 2021). There are numerous techniques and methods to apply ML models to business problems; for instance, clustering can be used to detect product errors (Zidek et al., 2016), quantitative evaluation (Onel et al., 2019) and equipment condition diagnosis (Rostami et al., 2016).

As computer power, sensor technology and available data increase by the minute, machine learning applications in the manufacturing industry are expected to grow at a fast rate (Columbus, 2020). As such, real-time data mining will play a vital role in the future, as organizations are now able to process, store and analyze increasingly high dimensional data like never before (Mussomeli et al., 2016), and applications based on them will substantially enhance manufacturing (Dogan and Birant, 2021).

2.2 Data Mining and Statistical Analysis

The ability to generate knowledge and effectively process information is undeniably recognized as a strategic asset. Data mining has thus become one of the most promising fields of machine learning, supporting data-driven decision making through the extraction of insights and pattern recognition. In recent years, data mining has been leveraged in complex manufacturing processes, augmenting quality diagnosis and quality improvement, thus becoming an emergent topic in the field of quality engineering (He et al., 2009).

It should be noted that statistical methods are the cornerstone of data mining and analytics, supporting the entire decision making process and analysis of results and insights brought forth by data mining techniques (Chen et al., 2018). In fact, the first step of the CRISP-DM methodology, disclosed in section 1.5, encompasses the task of exploring the data, which requires the use of statistical analysis to gather and summarize multiple data characteristics and highlight the most influential data (Ribeiro et al., 2017). Statistical methods are also required in the data preparation

phase to clean and construct the datasets, as well as in the evaluation phase, to analyze the project's results.

2.2.1 Statistical Indicators

An important and common topic in statistics is the analysis of variability, a crucial aspect in any manufacturing process that strongly impacts performance and expenditure. Variation is highly tied to production quality, as high variability leads to unreliable and unexpected outcomes, resulting in poor quality (Aba and Hayden, 2013). As such, most manufacturers aim to generate products or services with little to no variation, in order to maximize production quality and enhance customer satisfaction. Having said that, it is important to understand that there are two different types of variation affecting product quality: random and assignable variations. Random variations are caused by the natural characteristics of a manufacturing process and cannot be eliminated completely (He et al., 2009). On the other hand, assignable variations often result from a faulty manufacturing setup and can be traced back to the operator, the materials, machinery or the environment (Aba and Hayden, 2013). As such, assignable variations are predictable and can therefore be removed once they're identified; hence manufacturers should focus on their detection and subsequent elimination.

Gorunescu (2011) argues that without statistics, data mining would not exist, as classic statistical techniques enable the identification of relations between variables when there is insufficient information about them. Descriptive statistics, such as the mean, median and standard deviation, provide insights on the average values of each variable, central values and data dispersion, respectively. Paired with correlation analysis and visual representations such as histograms and boxplots, these statistical methods enable data scientists to understand how each variable operates, how the data is distributed and in what way the variables are related to each other. These techniques provide insights into the quality of the data, depicting the existence of missing values or outliers that negatively bias results. For instance, boxplots illustrate how the values in the data are spread out, explicitly identifying the existence of outliers. Histograms, on the other hand, allow analysts to get a sense of the variability of the statistical data. Moreover, a correlation matrix discloses how the variables are related to each other, where the correlation coefficient represents the relationship between two variables, which can be positive or negative.

2.2.2 Data Mining and Feature Selection

Statistics can also be used to depict how the sample data impacts the target variable. High dimensional data usually contains noise and irrelevant features, leading to the deterioration of the machine learning model's prediction capability and overall performance (D'Souza et al., 2020). Therefore, data scientists must be able to identify the most relevant features to predict the target variable to pose as inputs in the ML model. This not only enables the decrease of computational costs, but also leads to an improvement in the prediction accuracy (Haq et al., 2019).

According to Kuhn and Johnson (2019), one of the techniques used by data scientists to determine the best set of features that enables the construction of a reliable machine learning model is the stepwise regression algorithm, which can be a forward or backward selection. This approach relies on classic statistical metrics, such as the p-value or the R^2, to define which features should enter or leave the regression model. The coefficient of determination, the R^2, indicates the percentage of the response variable's variation explained by the model. Therefore, the higher the R^2, the better the model fits the data. A high p-value, on the other hand, indicates that the predictor variable does not impact the target variable and should not be included in the model. In a forward regression, there are no features in the selected set to begin with. The importance of each variable is tested and ranked by their individual ability to explain the variation of the target variable. If the resulting p-value is below the determined threshold, usually under 0.05, or the model's R^2 increases with the variable's inclusion, then the feature is considered relevant and is included in the model. In a backward regression, the model starts off with all the features included in the set, and if their removal increases the model's R^2 or their p-value exceeds the determined threshold, they are removed from the model.

Likewise, decision trees also rely on statistics to learn how to best split the dataset into smaller subsets to predict the target value. The "leaf" or node of the tree represents the condition or test, while the "branches" or edges represent the possible outcomes. The splitting process ends when no further gain can be made, or a preset rule is met, e.g. reaching the maximum depth of the tree. In a random forest algorithm, many individual trees are constructed to build the model. In this case, feature importance is calculated as the decrease in node impurity weighted by the probability of reaching that node. In regression models, the node impurity is obtained by calculating the variance reduction, using the Mean Square Error or the Mean Absolute Error, as shown in Table 2.1. For classification, the Gini impurity or the entropy are calculated to capture the node impurity. The node probability is calculated by dividing the number of samples that reach the node by the total number of samples. High values represent the most important features, enabling data scientists to understand which variables were more relevant to build the model (Ronaghan, 2018).

Table 2.1: Node impurity formulas

Impurity	Task	Formula	Description
Gini impurity	Classification	$\sum_{i=1}^{C} f_i(1-f_i)$	$f_i$is the frequency of label i at a node and C is the number of unique labels
Entropy	Classification	$\sum_{i=1}^{C} f_i log(f_i)$	$f_i$is the frequency of label i at a node and C is the number of unique labels
Variance / Mean Square Error (MSE)	Regression	$\frac{1}{N}\sum_{i=1}^{N}(y_i-\mu)^2$	y_i is the label for instance, N is the number of instances, and μ is the mean given by $\frac{1}{N}\sum_{i=1}^{N} y_i$
Variance / Mean Absolute Error (MAE)	Regression	$\frac{1}{N}\sum_{i=1}^{N}\lvert y_i-\mu\rvert$	y_i is the label for instance, N is the number of instances, and μ is the mean given by $\frac{1}{N}\sum_{i=1}^{N} y_i$

In short, the role of statistical methods in the construction of machine learning models is undeniable. On top of supporting the data understanding, preparation and evaluation phases, statistics can point the user towards the most important variables in a model, providing relevant insights for manufacturers.

2.3 Smart Manufacturing Use Cases

There is a potential to improve industrial operations through a number of identified data-driven smart manufacturing use cases, as they cover multiple industrial fields, such as transportation, manufacturing, healthcare, energy production, among many others (Vijayaraghavan and Leevinson, 2019). According to Tao et al. (2018), the most promising applications of Industry 4.0 for manufacturing processes are the following:

1. *Smart Design:* Manufacturers can no longer expect an increase in sales solely based on cost reduction and improved quality; their focus should be on products that fulfil individual preferences, hence the need for smart design (Wang et al., 2017). Involving users in the front end of the design process enables the co-creation of value, which is essential to achieve mass individualization and create long-term relationships (Pessoa and Becker, 2020). Robots, smartphones, and advanced image recognition can be used to collect powerful user data, such as consumers' behaviour, user-product interactions and consumer preferences, enabling the translation of the customer's voice into unique, desirable product features (Tao et al., 2018). Another example is the use of 3D printing in rapid prototyping, which enables faster user-feedback, as well as higher production flexibility (Pessoa and Becker, 2020).

2. *Material tracking and distribution:* Efficient material distribution requires the right material to be delivered to the right equipment at the right time (Tao et al., 2018). Technologies such as automated storage systems are being used to reduce the time of component searching and transportation by analyzing which parts are used more frequently to store them closer to the production line (Zin and Vogel-Heuser, 2019). It is also important to track materials and their condition to guarantee quality specifications (Tao et al., 2018). RFID-enabled positioning systems in AGVs use identification tags to track the real-time material condition, such as location, quality or status, enabling efficient material deliveries (Lu et al., 2017).

3. *Performance monitoring:* For operators to be able to quickly react to machine failures and performance issues, the manufacturing process, along with the whole equipment line and material environment, must be constantly monitored (Zin and Vogel-Heuser, 2019). Due to the predictive capabilities of big data analytics, alerts and recommendations can guide operators into time-efficient adjustments (Tao et al., 2018). Production disruptions on the shop floor, such as order tardiness, are usually caused by irregularities, namely equipment failure or lack of material. AI algorithms can be used to minimize these disruptions, as they capture patterns in time series from the collected information, ranging from material consumption data, energy consumption data, vibrations, to rotation rates (Tao et al., 2018).

4. *Quality Control:* The criteria used to accept or reject a product can range from its owns dimensions to the machine's pressure or temperature requirements (Ogorodnyk et al., 2021). This data concerning product quality and other parameters regarding geometry, location, tolerance and the machine, can be collected using different sensors, RFIDs and machine vision applications (Li et al., 2015). As a result, quality control can be deployed using big data analytics for early detection of quality defects and to perform quick diagnoses of their root causes (Tao et al., 2018).

5. *Smart planning and scheduling:* Before the product manufacture takes place, it is essential to plan and schedule the production process, taking into account the availability of resources and materials, as well as the production capacity of the manufacturing facility (Tao et al., 2018). Several issues that affect scheduling, such as volatile market demand trends, or workforce shortage, can be analyzed by big data applications and translated into effective decisions (Parente et al., 2020). Furthermore, cloud manufacturing can also be employed, for instance, to adjust to new product portfolios by increasing the flexibility of responsiveness to the changing customer demands (Erol and Sihn, 2017).

6. *Predictive maintenance:* Data analytics and machine learning algorithms can be used to provide a forecast of when equipment maintenance will be necessary, resulting in reduced unplanned downtime (Balamurugan et al., 2019). Predictive analysis combines the equipment's historical records with real-time data to analyze the tendency for deterioration, the remaining lifetime of components and the root cause of certain faults (Tao et al., 2018). As a result, manufacturers can employ precautionary maintenance to prolong equipment lifetime and reduce maintenance costs (Zhang et al., 2015).

The advances in Industry 4.0 technologies have profoundly impacted manufacturing processes, yielding numerous opportunities for business managers. Big data is empowering organizations to adopt data-driven strategies that enable agile responses to the ever-changing market conditions, as well as the creation of new, connected experiences for customers, partners, and the workforce. The IIoT paradigm is considered as one of the main trends affecting businesses today and in the future, as industries thrive on modernising systems and equipment to deal with disruptive technologies and keep up with the market volatility (Vijayaraghavan and Leevinson, 2019).

The present dissertation embeds the emerging Industry 4.0 and AI technologies in a business environment to optimize and enhance a particular manufacturing process in the oilseed segment. Powered by IIoT, multiple sensors are leveraged to exchange data and feed machine learning algorithms, where a selection of data mining techniques are applied to enhance asset performance monitoring and quality control. The projected solution enables the business under study to gradually shift from the traditional supply chain to a responsive, digital supply network, capitalizing on the predictive capabilities brought forth by AI technologies and statistical analysis to react to equipment and workforce breaches efficiently. The goal is to increase the final product's safety and quality indicators, provided the technical challenges and hurdles of a digital transformation are overthrown.

Chapter 3

Business Understanding

This chapter presents the scope of this dissertation and the business at hand, starting with a description of the oil extraction process. The focal point is the desolventizing process and its operational variables, as a significant part of the total solvent loss in the extraction occurs in this phase. An analysis of the current state of business affairs follows, with a subsequent proposition for future operations. Finally, the study's business goals are exposed accordingly.

3.1 Oilseed Processing

Oilseed processing aims to obtain high-quality oil while minimizing undesirable elements to achieve high extraction outputs and produce meals that respect international quality and safety regulations. Solvent extraction is the most effective technique used to recover oil from oilseeds. It involves bringing oilseeds in contact with a liquid solvent to obtain the dissolution of oil. The reason behind this method's popularity results from the high proportions of recovered oil, leaving behind only 0.5% to 0.7% residual oil in the raw material. The most widely used solvent for commodity vegetable oil extraction is hexane, currently used by the client at hand.

The solvent extraction technique is illustrated in Figure A.2 in Appendix A. It begins with the preparation of the raw material, required to guarantee that every oil-bearing cell is brought in contact with the solvent. The seeds are crushed in a corrugated roller mill, heated and then flaked between a pair of rolling mills to increase their surface area. The prepared material then enters the extractor, where a bed of solids is formed, shifting the soybean cake (solids and oil) and the miscella (hexane and oil) in opposite directions, resulting in a continuous counter-current extraction. This system enables the extraction of highly concentrated miscella for the upcoming distillation. The product that results from the oil extraction is called marc, which is a mixture of solids and hexane. The residual solvent is then removed in a desolventizer-toaster (DT) in three different sections for pre-desolventizing, desolventizing and toasting. Subsequently, the miscella is treated in a distillation operation, removing the hexane to form crude oil. This enables the recovery of hexane which is then reused in the extraction process.

The client is currently facing an operational issue, where, after desolventizing the extracted material in the desolventizer-toaster and dryer-cooler (DT/DC), the remaining solvent in meals is often above the legislated limits and exposes a high variability. As a result, a strong impact on industrial safety can be expected, as well as environmental contamination, product quality deterioration and escalated expenditure. A detailed analysis regarding the instability of the solvent extraction process in the DT/DC is therefore carried out in this dissertation.

3.2 DT/DC Equipment

The desolventizing, toasting, drying, and cooling procedures can be completed in a thermodynamic system, referred to as a DT/DC, which is shown in Figure A.3 in Appendix A. The equipment is structured as a single vessel, with multiple trays, where the desolventizer-toaster (DT) trays are in the top half, and the dryer-cooler (DC) trays are in the bottom half. The thermodynamic equilibrium between the liquid and vapor phases is achieved as the flour progresses through the different trays of the equipment, with the solvent being almost completely evaporated and most of the steam condensed.

The DT is a vertical, cylindrical vessel whose trays are heated using steam, hot water and oil. Its main purpose relies on separating hexane from the soybean oil meal, recovering as much solvent as possible, and producing high-quality meals with low energy consumption. There are two types of DT trays, the pre-desolventizing and desolventizing trays, designed with an upper and lower plate, as well as structural members in between, prepared to hold pressurized steam.

The process in the DT/DC initiates when the material arrives from the extractor and enters at the top into the 1st to 3rd floors of the DT. These are the pre-desolventizing trays, where the material is heated through indirect steam. The sole purpose of these trays is to provide conductive heat transfer through their upper surface to the material filled with solvent, which is supported above. The material is mixed above each tray and transferred downward from tray to tray, through agitating propellers, anchored to a central rotating shaft.

The desolventizing task then takes place in the five central trays, the 4th to 8th floors. At this point, in addition to the indirect heating, these trays provide direct heating that results from direct steam passing through the bed of solids. Sluice valves are used to maintain the flow of material between the trays, allowing solids to pass through according to their level. These trays have a dual purpose, first to provide consistent direct steam into the meal layer, and second, to provide conductive heat transfer through its upper surface to the wet material supported above. After being desolventized and toasted, the soybean meal should exit the DT with a hexane residual lower than 500 ppm, temperature between 105 and 115°C, and humidity content around 20%.

Subsequently, the material enters the DC, that much like the DT, has upper and lower plates; however, in this case, the structural members in between are designed to distribute low-pressure air vertically into the meal layer supported above. There are also two types of DC trays: steam drying trays and air cooling trays, and these are located in the 9th to 11th, and 12th floor, respectively.

The three drying trays are designed to evenly introduce hot air into the meal. They are also drilled in order to allow the passage of air through the bed. In this case, the material is also transferred between floors through sluice valves. Finally, in the last tray, the 12th floor, a blower injects cold air into the chamber, in order to cool the meal before it leaves the DT/DC for storage.

3.3 Operational Parameters

Determining the optimum DT/DC configuration for oilseed processing is a rather complex task that relies on multiple factors. The most influential parameters in the different floors of the equipment are the following:

- Steam/floor temperature;
- Direct steam flow;
- Level (height) of solid material;
- Discharge speed of the valves;
- Driving force of the blowers;
- Steam/floor pressure.

It is important to control the steam temperature to enhance the equipment's efficiency. Since hexane presents a low boiling point, around 67-70°C, most modern DTs are operated with temperatures ranging from 70 to 75°C, in order to maintain the low solvent loss, while assuring a safety margin. If the temperature is below this minimum threshold, the overall efficiency can be tarnished, as the amount of evaporated hexane will be reduced, resulting in wasted solvent and possibly a waste of the final product. As a result, it is important to maintain the steam temperature as low as possible, minimizing the total energy consumption, while preserving a reasonable level of evaporated solvent.

Regarding the addition of direct steam, its purpose is to regulate and secure a constant temperature at the top of the DT, while controlling the temperature gradient along the desolventization trays. It is important to maintain a sufficiently high direct steam flow rate per unit area, in order to guarantee an adequate amount of desolventized meal in the DT, as the residual hexane in the solids decreases with increasing vapor density.

Furthermore, the level control of the solid material is also a crucial factor to consider, especially in factories that process different seeds with different daily flows. This can cause inefficiencies in the desolventizer, due to the seed's distinct properties, resulting in the need to adjust the solid levels in the trays. The discharge speed of the sluice valves is thus regulated for this purpose. The amount of hexane present in the flour exiting the desolventizer is a good indicator of whether the process suffered any concerning issues.

Finally, the centrifugal blowers are used to pressurize the air inside the DC and drive cool air through the solid material. As the cool air enters the trays, it flows upward through the meal, and

its speed must be regulated to create a partially fluid meal. As such, it is important to control and maintain the appropriate driving force of the blowers to ensure the right pressure and temperature in all the floors of the equipment.

The DT/DC is a thermodynamic system in equilibrium, meaning that parameters such as the temperature, flour levels, steam and pressure all interact and influence the final product's solvent concentration. As such, these variables must be carefully monitored and controlled on each floor. An essential factor to consider is the definition of the process's setpoints, as these represent the desired or target value of the flour levels, temperature, steam and pressure in the equipment. Setpoints are directly handled and manually defined by the operator, who sets them according to his acquired knowledge and the process's state of the art.

3.4 Quality Management

Having understood the technical aspects, it is important to proceed to an analysis of the ongoing operations occurring on the shop floor. Evaluating the current state of the business' processes provides an insight into the scope and origin of their affairs, driving valuable opportunities for improvement and quality management. An as-is analysis concerning the company's solvent extraction technique was therefore conducted, followed by a to-be analysis, exposing the proposed strategy for the future.

3.4.1 Current State Analysis

The main problem presently faced by the client is the inconsistency and high variability of the hexane concentration present in the flour leaving the DT/DC equipment. Considering that a concentration of 500 ppm or under is required to meet the legislated limits, it is important that the client is able to rigorously control this output.

Currently, the production process begins with an assessment and planning of the monthly needs, as shown in step 1 in Figure 3.1. This analysis is performed by the production director, who subsequently releases the respective production orders. In step 2, the supervisor receives and plans the orders on a daily basis and releases them to the operator, who then executes the given tasks in step 3. The process monitoring is then carried out in the 4^{th} step, by both the operator and the supervisor, who are able to monitor the operational parameters in the SCADA control system. Finally, the supervisor confirms the production order was successfully executed and whether the DT/DC is operating accordingly, before the whole process restarts.

Concurrently, once a week, the production director requests a laboratory analysis of the flour's hexane concentration, as shown in step 1.1. Then in step 1.2, the operator collects a product sample and delivers it to the laboratory, where numerous analyses for the entire factory are carried out (step 1.3). Since this particular analysis is not a priority, the results are obtained only eight hours later, causing a serious delay. This holdup forecloses the opportunity to act on root causes in real-time; when a problem in the equipment is diagnosed, the required adjustments will be made with a significant delay. Moreover, further analysis must be carried out to pinpoint the effects of

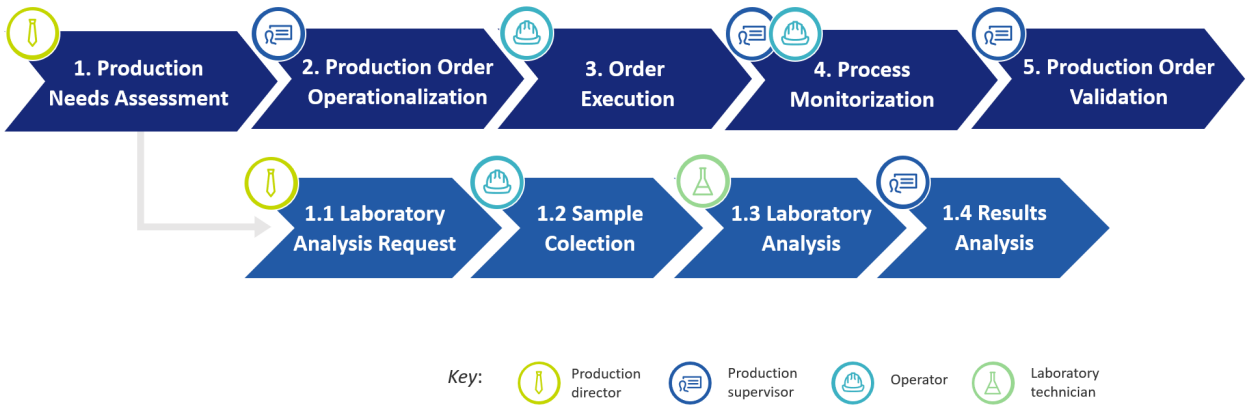

Figure 3.1: As-Is Process

such adjustments, which will also be delayed. When the supervisor receives the deferred results, he analyzes them and releases adjustment orders (step 1.4) that will also be overdue.

Every holdup and delay not only leads to increased expenditure but also results in a substantial waste of the final product. Faced with a volume-driven business model, the wasted batches of flour represent a considerable loss for the company. Furthermore, the solvent extraction process represents over 50% of the factory's energetic expenses and the optimization of the operations involved could result in a significant cost reduction for the future.

3.4.2 Future State Proposal

The proposed future process aims to increase the agility and efficiency of the whole operation. The projected solution discloses the use of a cloud data visualization tool, SAP Analytics Cloud (SAC), that enables business users to create interactive dashboards, supporting the last mile of the decision-making process, with its augmented analytics capabilities, powered by AI and ML. Furthermore, the solution integrates the developed predictive models, providing the different hierarchical levels in the business with an opportunity to monitor and control the operational parameters in real-time, and perform the necessary adjustments in due course.

The 1st step of the proposed journey is identical to the one referred earlier in the as-is analysis. In the 2nd step, however, the supervisor is able to monitor the process parameters with the support of the dashboards that embed the results from the analytical models. This way, the supervisor is able to depict the optimal operational parameters, analyze the current problems in the process, and create adjustment orders to address them. The operator then executes the respective tasks in step 3, as mentioned earlier. Subsequently, in step 4, the process monitoring is performed with the support of the designed dashboards, not the SCADA system. Finally, in the 5th step, the supervisor uses the proposed dashboards to validate the production order and confirm whether the equipment is operating smoothly. Regarding the parallel laboratory analysis steps, the only modification required is in step 1.4, where besides analyzing the results and releasing process corrections, the supervisor must upload the results onto the technological solution for future assessment.

Consequently, the proposed transformation allows the client to monitor the hexane concentration in real-time, as opposed to the current situation, where results are obtained with a significant delay. The ability to act in real-time and adjust the necessary parameters in due course grants the client the fundamental agility required to increase their quality standards.

3.5 Business Goals

The present dissertation project aims to improve the outlined process by enhancing specific actions undertaken in the client's current operations. Regarding the objectives mentioned in section 1.4, the business goals were divided into three distinct levels and are outlined below:

1. Obtain a real time estimation of the flour's ultimate quality:
    - Develop a machine learning model to provide a real-time prediction of the final product's quality based on the input variables;
    - Depict the variables that provoke the most impact and provide the best explanation for the concentration of solvent (hexane) in the product, as it leaves the equipment.
2. Optimize the definition of input parameters:
    - Define the optimal values of each input variable to depict their ideal setpoints, according to the specified hexane concentration that agrees with the quality and safety indicators (500 ppm);
    - Segment the data into different groups based on distinct ranges of hexane concentration, depicting the maximum and minimum value of each input parameter for each range of values. This provides an opportunity to monitor the process with better insights on the parameters that should be adjusted.
3. Enhance process calibration and supervision in real-time:
    - Leverage a technological solution, SAP Analytics Cloud, to implement the developed predictive models and display the information in visual, interactive dashboards;
    - Monitor and control each operational parameter, analyzing their trends and variability over time, comparing them to the defined setpoints;
    - Visualize the optimum thresholds defined for each parameter and compare them to their current values in real-time;
    - Access information regarding past laboratory analyses and their timestamps, aiding the management of future requests.

This analysis concludes the first phase of the CRISP-DM methodology, exposing a thorough understanding of the business, the terminology used and the listed objectives.

Chapter 4

Model Development

This chapter describes the data provided by the client, followed by a thorough analysis of its quality and rising insights. Then, all the steps performed in the data preparation phase are exposed, including the data aggregation, outlier treatment, dataset construction, and correlation analysis. Subsequently, several data mining techniques are selected and tested on the different datasets. Finally, an assessment concerning the models' end results is carried out, as well as an evaluation of their ability to meet the outlined goals.

4.1 Data Understanding

To begin with, it is important to understand the provided data and its alignment with the business problem. The information disclosed by the client derived from sensors installed throughout the extraction process, as well as manual recordings of the flour's hexane concentration, analyzed in a laboratory. The date and time of each variable are also stored and disclosed accordingly.

4.1.1 Data Presentation

There are over one hundred sensors in the factory's oil extraction process that record data every 5 seconds, displaying the information in a supervisory control and data acquisition system called SCADA. A schematic representation of the client's SCADA is exposed in Figure A.4 in Appendix A. The most interesting variables were selected, alongside the client, resulting in 42 operational parameters to be analyzed.

The sensors under study and the applied abbreviations are the following:

- Speed controller (SC);
- Level indicator transmitter (LIT);
- Temperature transmitter (TT);
- Pressure transmitter (PT);

- Intensity transmitter (IT).

The objects monitored by these sensors and their respective abbreviations are the following:

- Blowers (fan);
- Sluice valves (valv);
- Flour (prod);
- Flour leaving the equipment (prodOut);
- Vapor leaving the equipment (gasOut));
- Floor of the DT/DC (floor);
- Lubrication pump (pump);
- Main engine (motor).

A systematic nomenclature was defined for all the variables, whose structure was defined as follows: "Type.Of.Sensor_Monitored.Object_Floor.Number". The monitored floors of the DT/DC equipment are the 1st floor and all those between the 4th and 12th floor. These operating parameters are summarized in Table 4.1, disclosing which variables are directly controlled by the operator, the floors that each sensor monitors, the measurement unit, and a short description regarding each variable.

Table 4.1: Summary of the DT/DC operating parameters

Variable	Directly Controlled?	Floor (X)	Unit	Description	Total
SC_fan_X	Yes	X ε (9;10;11)	Hz	Driving force of the blowers	3
SC_valv_X	Yes	X ε (4;5;6;7;8;12)	Hz	Sluice valves speed controller	6
LIT_prod_X	Yes	X ε (4;5;6;7;8;12)	°gr	Flour level	6
TT_gasOut_X	No	X ε (1)	°C	Output vapor temperature	1
PT_gasOut_X	No	X ε (1)	mmH_2O	Output vapor pressure	1
TT_prod_X	No	X ε (1;4;5;6;7;8;9, 10, 11, 12)	°C	Flour temperature	10
TT_floor_X	No	X ε (9;10;11)	°C	Floor temperature	3
PT_floor_X	No	X ε (4;5;6;7;8;9, 10, 11, 12)	mmH_2O	Floor pressure	9
TT_prodOut_X	No	X ε (12)	°C	Output flour temperature	1
TT_pump	No		°C	Pump temperature	1
IT_motor	No		A	Main engine current intensity	1

The client also provided data regarding a selected number of setpoints. This enables the analysis of the equipment's responsiveness to the benchmarks and target values defined for each operational parameter. Table 4.2 summarizes the available setpoints, which follow a similar nomenclature structure as the homologous variable: "SP_Type.Of.Sensor_Monitored.Object _Floor.Number".

Lastly, laboratory analyses are performed to obtain information regarding the hexane concentration present in the flour. The resulting data is summarized in Table 4.3, which includes records

Table 4.2: Summary of the setpoint variables

Variable	Floor (X)	Unit	Description	Total
SP_TT_gasOut_X	X ε (1)	mmH_2O	Setpoint of the vapor temperature leaving the DT/DC	1
SP_LIT_prod_X	X ε (5;6;7;8)	°gr	Setpoint of the flour level	4
SP_TT_floor_X	X ε (9;10;11)	°C	Sepoint of the floor temperature	3

regarding the hexane concentration, as well as the seed's origin and type of flour. The information resulting from these analyses aims to provide the client control over the equipment's output, enabling the quality control of the final product. However, the product samples are usually collected by the operator only once a week, and the reports are obtained eight hours later. As such, the results arrive with a significant delay, preventing a reliable and real-time fine-tuning of the process.

Table 4.3: Summary of the laboratory measurements

Variable	**Unit**	**Description**
Hexane concentration	ppm	Extracted flour hexane concentration
Origin	USA or Brasil	Where the soy seeds come from
Type	44 or 47,5	Type of flour in process

In short, the data provided by the client includes information regarding 42 operational parameters, 8 setpoints and manual laboratory measurements that depict the solvent concentration of the final product, providing insights on its quality.

4.1.2 Exploratory Data Analysis

In this phase, primary research is performed on the data, aiming to discover patterns, signs of anomalies, and data quality issues. Summary statistics and graphical representations were carried out to help build familiarity with the data, providing early insights that prepare the ground for the data preparation phase.

To begin with, a basic statistical analysis was performed on each variable, calculating their mean, standard deviation, and minimum and maximum values over an hour. Table 4.4 shows the statistical summary of a selection of variables. These results demonstrate that most operational parameters present a variation of approximately zero over the course of one hour. The standard deviation of variables *SC_fan_9*, *TT_prod_5,* and *PT_floor_8* is in fact zero, meaning that their values remain constant in an hour. This outcome is expected due to the thermodynamic equilibrium present in the DT/DC equipment. As such, an opportunity arises to improve the quality of the data and reduce its dimensionality through an hourly aggregation. However, this must first be analyzed with the client to confirm that this variation is residual.

Due to further insights from Table 4.4, a decision was made, alongside the client, to remove the variable measuring the flour's temperature on the 4th floor (*TT_prod_4*). This resulted from the fact that this parameter presents a variation between 62 and 63°C, which the client confirmed was

Table 4.4: Statistical summary of a selected number of variables

	SC_fan_9	SC_valv_12	LIT_prod_6	TT_prod_5	TT_prod_4	TT_floor_11	PT_floor_8
Mean	37	18,35	44,27	105	62,89	24,05	53
Std	0	2,63	2,77	0	0,31	0,23	0
Min	37	13	38	105	62	24	53
Max	37	24	52	105	63	25	53

a misleading observation, since the actual temperature of the floor should be the same as on floors 5, 6, and 7. The cause of this misreading is the location of the sensor; positioned at the top of the floor and faced with a usually low level of flour, the sensor is rarely submerged in the product. Therefore, the recorded temperature is inaccurate and could negatively bias future analysis, hence the decision to exclude it.

Subsequently, visual analyses were performed on each variable through histograms and boxplots. This study aimed to grasp a better understanding of the distribution of each variable, while depicting the existence of outliers. The graphical representations of a selection of variables are demonstrated in Figure 4.1. The constructed boxplots indicate that all the operating parameters detain a high number of outliers, yet their removal or replacement must be clarified with the client, as they could derive from machine breakdowns, maintenance interventions, among other known reasons. In fact, further analysis revealed that the majority of the variables present a high number of zeros, which could be explained by a production shutdown or sensor failures, in which case they could be treated as outliers. In addition, Figure 4.1 demonstrates that the exposed variables present an approximately normal distribution, as did the other variables that are not represented.

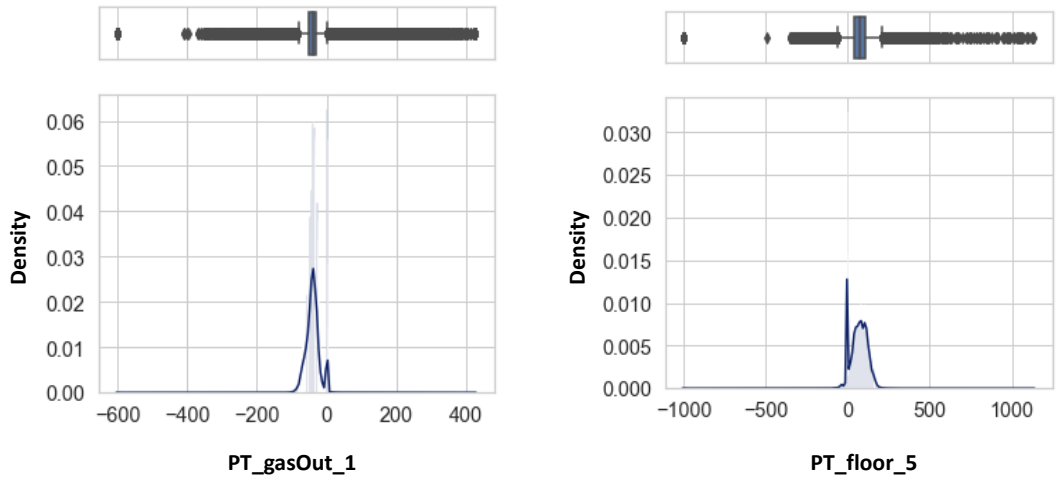

Figure 4.1: Histograms and boxplots for *PT_gasOut_1* and *PT_floor_5*

A time-series analysis was then conducted for each variable in order to examine their performance over time. Two diagrams are exposed in Figure 4.2, illustrating only the month of March for better visualization. These graphs corroborate the fact that the operational parameters detain a high number of zeros and that these occur simultaneously. It should be noted that this analysis was carried out for all the available months and variables.

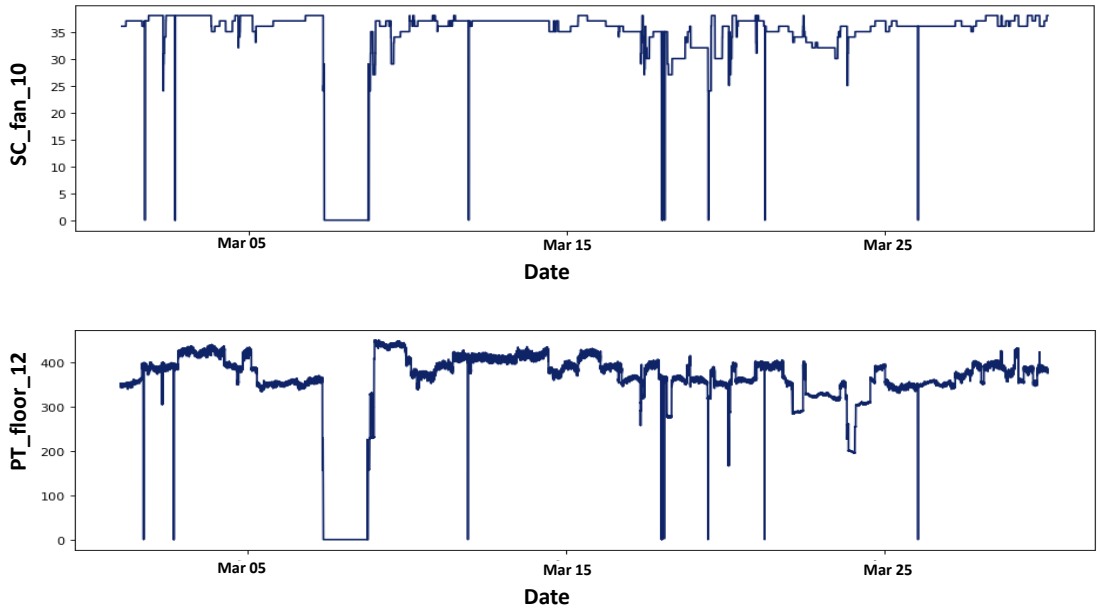

Figure 4.2: Time-series analysis of *SC_fan_10* and *PT_floor_12*

A similar analysis was conducted for each setpoint and the corresponding operational parameter. Figure 4.3 shows how the temperature on the 9th floor of the equipment responds to the setpoint defined by the operator. The graph illustrates how the actual values fail to reach the established target, resulting from the fact the implemented setpoints are calculated on theoretical, state of the art values that the operators define based on their acquired knowledge of the process. As a result, this analysis reveals the need to define the optimum setpoints that the equipment is in fact able to achieve.

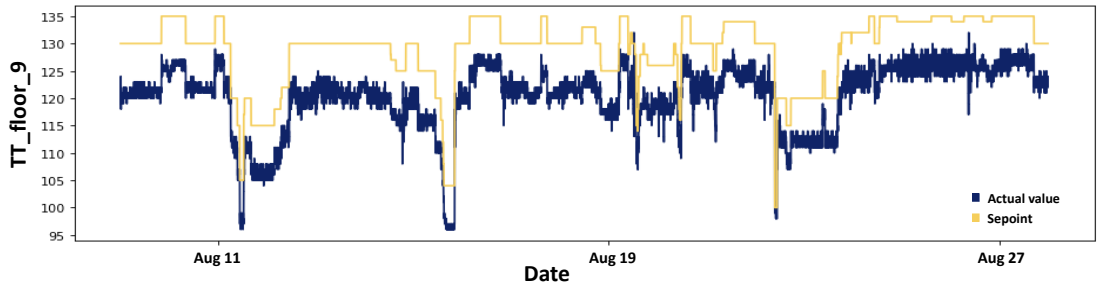

Figure 4.3: Time-series analysis of *TT_floor_9* and corresponding setpoint

Furthermore, it is important to understand the correlation between all the operational parameters to enable the dimensionality reduction of the data. This is achieved by removing variables that are highly correlated with each, since they add no value to the analysis. The constructed correlation matrix is represented in Figure B.1 in Appendix B, where it appears that all the variables are strongly correlated. However, this analysis is misleading due to the high number of outliers present in all the parameters. These must be removed, before a new correlation analysis is performed subsequently.

Finally, with regards to the laboratory measurements, it is important to note that the number of available records is small, as the client only provided 66 observations, whereas the operational parameters detain 4.589.632 records. As a result, an opportunity to improve the quality of the data in the future has been identified. The laboratory missing values can be estimated before merging these records with the operational parameters to ensure an enriched dataset.

The insights resulting from this exploratory data analysis can be summarized as follows:

- There is no significant variation in the DT/DC operational parameters over the course of one hour, in procedural terms;
- All the operating parameters detain a high number of outliers;
- There is a very short number of laboratory measurements when compared to operational records.

4.2 Data Preparation

In a predictive modelling project, raw data must be pre-processed before being used to fit and evaluate a machine learning model. The data preparation phase enables the transformation of raw data into a suitable structure by correcting errors and statistical noise, identifying the most relevant input variables, and creating compact data projections.

4.2.1 Data Aggregation

Through the statistical analysis exposed earlier in section 4.1.2 and the information provided by the client, it is possible to conclude that the operating parameters of the DT/DC equipment show no significant variation over an hour. As a result, and due to the strong inertia inherent to the process, the hourly aggregation of the data was considered an adequate approach that averts disturbing the results, while reducing the data dimensionality.

When performing the data aggregation, it is important to select the best metric for the case at hand. The mean, median, and mode of each hour were considered for this purpose. However, it should be noted that due to the parameters' high number of outliers identified earlier, both the mean and mode would misrepresent the results, since these metrics are highly sensitive to outliers. On the other hand, the median takes into account all the values from the period of an hour and selects the central value, leading to higher accuracy.

The data was thus aggregated using the median, resulting in a dataset with 18.169 records, which accounts for a dimensionality reduction of 99,6%. This enables the optimization of the analysis in terms of CPU processing, while improving the proportion of available laboratory values and operating parameters records.

4.2.2 Outlier Treatment

Within the scope of data mining, outlier detection seeks to uncover patterns that diverge from the expected behaviour, accounted by individual records, distant from the remaining observations. It is important to understand the reason behind their occurrence in the business context, in order to guarantee they are dealt with properly.

As such, and after rigorous examination, the client confirmed that the observed zeros, exposed in section 4.1.2, took place during a production break at the factory. These observations are thereby considered as outliers, and their removal is thus required to prevent a negative bias of the results.

4.2.3 Dataset Construction

After removing the outliers, the operational records and the laboratory measurements were merged together, resulting in the construction of three datasets. The first dataset contains the original 66 available laboratory records and the respective operational parameters. This means that the operational data is missing 4.589.566 data points, which accounts for 99,9% of missing values. Therefore, and considering that the client confirmed that the hexane concentration presents close to no signs of variation during the course of one hour, a second dataset was constructed, where the laboratory records were duplicated for the previous and succeeding hour of each measurement.

Furthermore, a third dataset was developed, using the KNN algorithm to depict the missing values of the hexane concentration. This method relies on the other variables present in the dataset to replace the missing records, classifying data points based on their similarity.

An analysis was carried out to assess the optimum value of k for the KNN algorithm. Table 4.5 shows the mean and standard deviation of the hexane concentration for each value of k, where the last column displays the statistics of the actual laboratory measurements. The study demonstrates that the mean values are quite similar in each approach, whereas the standard deviation shows a wider variation. Usually, the lowest standard deviation would be preferable, since it represents a narrower spread of values. However, in this context, the chosen value was k=3, since this approach grants the mean and standard deviation closest to actual data provided by the laboratory measurements.

Table 4.5: Comparison between the different k parameters for the KNN method

	k=3	**k=5**	**k=7**	**k=9**	**Actual Laboratory Data**
Mean	768,75	778,85	769,08	791,20	785,34
Standard Deviation	172,98	144,13	166,40	109,44	276,705

In the end, the constructed datasets were the following:

1. Dataset A - The original laboratory measurements (66 observations);
2. Dataset B - The laboratory measurements duplicated for the N^{th+1} and N^{th-1} observations (198 observations);
3. Dataset C - An estimation of the missing laboratory values using the KNN algorithm (5.788 observations).

4.2.4 Correlation Analysis

A correlation matrix was then constructed for each dataset mentioned above. This analysis is no longer biased by outliers since the data has been cleaned accordingly.

The resulting correlation matrices for datasets A, B and C, are represented in Appendix B in Figures B.2, B.3, B.4, respectively. Table B.3 in Appendix B summarizes which variables presented a high correlation, and as a result, which ones were removed. Variables with a high correlation, over 0.85, were removed, provided they're not affected by direct human control. For instance, in dataset A, variables *SC_fan_10* and *SC_fan_11* are 92% correlated, yet neither were removed since the operator has direct control over them, and removing them can compromise the model's scalability. Ultimately, five variables were removed from datasets A and B, and eleven were removed from dataset C, resulting in 37 input variables for models using datasets A and B, and 31 for dataset C.

4.3 Modelling

The modelling phase of the CRISP-DM approach involves building and assessing various models based on several data mining algorithms. There are four main tasks involved in this stage, starting with the selection of the modelling techniques, followed by the partitioning of the data into training and test subsets, the model construction, and finally, its assessment.

4.3.1 Selection of the Modelling Technique

Considering the nature of the project, the matter is categorized as a supervised machine learning problem, since it aims to predict a target variable, Y, from 42 known operational parameters. Provided the goal was to simply estimate whether the final hexane concentration was within or outside the legislated limits, the matter could be treated as a classification problem. However, regarding the goals outlined in section 3.5, this dissertation proposes to encounter knowledge on the hexane concentration in real-time. The client requires this information to efficiently control and monitor the process on a daily basis. As a result, the matter at hand was considered a regression problem, and the data mining techniques selected to construct the predictive model were the following:

- Stepwise Regression;
- Random Forest;

- Multiple Linear Regression;
- Sliding Window.

The stepwise regression algorithm was selected due to its ability to manage a large number of potential predictor variables. It performs a model fine-tuning by choosing the best predictor variables from the available options. The random forest was considered since it is also highly efficient when dealing with a great number of features in the data, and when handling linear and non-linear relationships. This algorithm aggregates the output of multiple regression trees and is therefore expected to generate more accurate results than a simple regression tree. The multiple linear regression technique was selected to assess whether the dependent variable, the hexane concentration, can be estimated from the set of independent operational variables.

Finally, the sliding window technique was chosen to support the prediction of the target variable based on historical data of the independent variables. In this case, a time-series dataset is constructed, containing information of the operational parameters from the past X hours, where X is the window size, that will be optimized to obtain the smallest possible error. For this purpose, the datasets demonstrated in section 4.2.3 are not suitable, as they lack the historical data in between the available laboratory measurements. As such, the sliding window technique generates a new dataset that includes the actual values of the laboratory measurements and contains X-1 additional columns for each variable, comprising the data from the past hours. Figure 4.4 illustrates an example where the size of the window, X, is equal to 4 hours. The random forest and multiple linear regression are then applied to this dataset.

Δ t	SC_fan_11 (1)	SC_fan_11 (2)	SC_fan_11 (3)	SC_fan_11 (4)	SC_valv_4 (1)	SC_valv_4 (2)	SC_valv_4 (3)	SC_valv_4 (4)	...	Y
0	35	35	35	37	32	31	31	32	...	1310,1
1	35	35	37	37	31	31	32	32	...	844,5
2	35	37	37	37	31	32	32	29	...	434,3
3	37	37	37	37	32	32	29	27	...	562,9

Figure 4.4: Sliding Window dataset example

4.3.2 Test Design

A machine learning model aims to predict a target variable based on previously unseen data. Hence, the dataset must be split into training and test sets, before the model is built, ensuring that the model can be evaluated in an unbiased manner, using data it has never seen before.

The training dataset is the sample of data used to train and fit the algorithm. The model sees part of the data and learns from it, to perform its task at a high level of accuracy. The test set, on the other hand, is held back from the training dataset and is used to assess the model's accuracy against its target.

There are two main issues to be considered when splitting the data into the different subsets. Firstly, the training data must be large enough to allow the machine learning model to make predictions; it usually takes up at least 70 to 85% of the whole data. Secondly, in order to obtain more accurate results, the data must be partitioned in a balanced manner, meaning that the distribution of the target variable is approximately the same in each subset. In this case, a new attribute was generated for the hexane concentration, classifying each observation into one of 5 groups, based on their value:

- Group 1: $\leq 500ppm$
- Group 2: $]500ppm, 700ppm]$
- Group 3: $]700ppm, 900ppm]$
- Group 4: $]900ppm, 1100ppm]$
- Group 5: $> 1100ppm$

The datasets were thus split randomly, in a balanced manner, taking into account the group each observation belongs to, with 75% of data being used for training and 25% for testing.

However, this approach is unsuitable for the sliding window since the technique involves a time-series dataset that must be kept in order and cannot be split randomly. As such, a time-based cross-validation splitting method was applied for the sliding window to provide a statistically robust model evaluation. This method begins with a small data training subset, followed by the calculation of the respective prediction and accuracy. The test subset is then included as part of the next training subset, and the next data points are forecasted. The training subset increases until all the data has been tested. The forecasting accuracy is then calculated as the average of the validation metrics calculated in each test set.

4.3.3 Model Construction

Using the four data mining techniques identified above, the different algorithms were computed to predict the target variable - the hexane concentration.

To begin with, the stepwise regression technique with a forward selection was employed, starting with an empty model. The variables that provided the greatest statistically significant improvement of the model fit were then added one by one, until no further improvement occurred. The criteria used to determine which variable should be added was the lowest p-value. The operational parameters that proved to be statistically significant for this model, for each of the three datasets, are presented in Table B.1 in Appendix B, along with the corresponding p-values. The variable *IT_motor* proved to be one of the most statistically significant in all three datasets, presenting a p-value of 0,00.

The backward selection technique was then computed accordingly. The model began with all the candidate variables, followed by a test on the elimination of each variable. Those whose loss provided the most statistically significant improvement of the model fit were removed, and the

process was repeated until no further variables could be removed, without a statistically significant deterioration of the model fit. The independent variables that proved to be statistically significant with the backward selection, for each dataset, are presented in Table B.2 in Appendix B, along with the corresponding p-values. It should be noted that before running the model for both stepwise regression techniques, the variables earlier removed in the correlation analysis were re-inserted, since the algorithm itself already excludes the variables that add no value to the model.

Subsequently, the random forest algorithm was constructed. This technique builds multiple decision trees and merges them together to obtain higher accuracy and robust predictions. There are two main steps involved in building a decision tree. First, the variables are divided into a set of distinct and non-overlapping regions. Then, a prediction is computed for each observation in each region, which is usually the mean value in the training set in that particular region. The predictions obtained from the random forest are the average of the predictions produced by the trees in the forest.

Next, the multiple linear regression model was computed. The algorithm calculates the minimum distance between each variable and an ideal hyper-plane, enabling the use of several explanatory variables to predict the outcome of a response variable, in this case, the hexane concentration. For this particular technique, the highly correlated dependent variables, exposed in Table B.3, that were not removed, were now excluded from the analysis to avoid over-fitting, since one of the assumptions of this technique is that the independent variables are not highly correlated with each other.

Finally, the sliding window technique was applied. A new correlation analysis was deployed since a new dataset is required for this technique. Similarly to dataset B, the new dataset contains 198 laboratory observations, yet it also includes the operational parameters' values for the missing hexane concentration observations. This enables the construction of a sliding window. The resulting correlated variables are presented in Table B.4 in Appendix B, where a total of 8 variables were removed. Furthermore, different sized windows were tested to calculate the resulting prediction accuracy, enabling the determination of the best-sized window. The results showed that a window of 6 hours leads to the smallest error. As such, the resulting model contained information of the past 6 hours of each operational parameter and 198 laboratory observations to compute its prediction. The random forest and multiple linear regression techniques were then applied to forecast the target variable.

4.3.4 Model Assessment

It is important to assess the model according to the data mining success criteria and the test design defined earlier. Since the current analysis is based on a regression problem, the four metrics used to evaluate the machine learning models were the following:

- **Root mean square error (RMSE)**: represents the sample standard deviation of the differences between predicted values and observed values (called residuals). It estimates how wide the residuals are dispersed;

- **Mean absolute error (MAE)**: represents the average of the absolute difference between the predicted values and observed value;
- **Mean absolute percentage error (MAPE)** : represents the average of the absolute percentage errors of the prediction. This metric presents a very intuitive interpretation of the prediction accuracy;
- **Coefficient of determination (R^2)**: measures how much variation in an outcome can be explained by the variation in the independent variables.

The results obtained using these metrics are represented in Table 4.6. It should be noted that some techniques result in a negative R^2, meaning that the computed model represents a weaker fit than a hyper-plane depicting the mean value. The R^2 is calculated as described in Equation 4.1, where RSS represents the Residual Square of Errors, computing the difference between the predicted and actual values, and TSS represents the Total Sum of Squares, calculating the difference between the actual values and the overall mean. As such, when the RSS is greater than the TSS, the R^2 is negative, indicating that the predicted values represent a poorer approximation than the overall mean.

$$R^2 = 1 - \frac{RSS}{TSS} \tag{4.1}$$

As shown in Table 4.6, the technique presenting the best performance is the random forest, with the highest R^2 and the lowest prediction error for datasets B and C. The results from dataset B reveal that 84% of the data fits the random forest model, and the predictions are obtained with an error of 14,82%.

Table 4.6: Validation metrics applied to each model for each dataset

Dataset	**Technique**	**Metric**			
		R^2	**RMSE**	**MAE**	**MAPE**
A	**SW B**	0,77	134,66	99,05	20,53%
	SW F	0,59	178,48	124,47	25,99%
	RF	0,04	238,24	191,08	28,38%
	MLR	-13,44	923,44	694,97	94,10%
B	**SW B**	0,56	180,22	136,98	21,26%
	SW F	0,62	168,22	124,10	20,19%
	RF	0,84	120,06	97,22	14,82%
	MLR	0,62	186,02	142,87	22,12%
C	**SW B**	0,41	133,52	101,17	13,74%
	SW F	0,42	132,11	99,62	13,51%
	RF	0,77	81,98	49,72	6,56%
	MLR	0,40	134,85	102,21	13,85%
Sliding Window	**RF**	-0,77	237,85	185,25	23,76%
	MLR	-7,39	531,68	383,92	52,83%

On the other hand, the predictions calculated with dataset C present a much smaller error, of 6,56%. This dataset contains over 95% more hexane concentration observations than datasets A and B, since the KNN algorithm was employed to estimate the missing values. However, these missing values were predicted based on a very small number of neighbours; hence the results are likely to be distorted. Furthermore, the analytical model computed its predictions based on these previously estimated data points. As such, the real values of the hexane concentration are not used to train and fit the model as it is for datasets A and B. For this reason, the errors for dataset C were found misleading and were not considered for the deployment phase.

In addition, the results for the stepwise regression proved to be lower than the random forest in dataset A. However, in dataset B, due to the rise in data observations, the random forest presents a vibrant improvement, with a 14% decrease of the prediction error and an increase of 80% for the R^2, whereas the stepwise regression presents a much smaller improvement. This hypothesis is reinforced with the multiple linear regression, where the error decreases from 94% to 22% from dataset A to B, and the R^2 improves from a negative value to 62%. The model's high prediction error and negative R^2 in dataset A is fairly expected since the algorithm uses only 66 observations to compute the minimum distance between each of the 37 independent variables and a hyperplane. The fact remains that, despite the misleading results from dataset C, the higher the number of observations, the lower the model's prediction errors.

Finally, with regards to the sliding window technique, the R^2 is negative for both the random forest and the multiple linear regression, meaning that this technique is not suitable for the available data. This outcome is also fairly reasonable, since there are 204 variables (6 hours x 34 variables) being used to predict the target variable, from only 198 past observations. The errors are smaller than those from dataset A, since there are more laboratory observations, but they are still significantly higher than dataset B. The high dimensionality of the dataset might be causing the overfitting of the model, which explains the higher prediction errors. However, there is a potential for the sliding window to become effective in the future, provided a great deal more observations are provided.

In short, there are two main conclusions that arise from this assessment. The first one is that the random forest is the best fit for this analysis, with dataset B providing the most accurate results. Secondly, the prediction errors show a clear sign of improvement as the number of laboratory observations increases. This proves that machine learning models perform better when more data is available to learn from.

4.3.5 Feature Importance

As mentioned in the literature review in section 2.2.2, there are many ways to determine the most important features in a model. One of the techniques is the stepwise regression that typically uses a p-value of 0,05 as a threshold to determine whether or not each variable has an impact on the hexane concentration. However, this particular feature ranking fails to take into account the possible interactions between the variables. Since there are 42 operational parameters in a thermodynamic system in equilibrium, it is important to consider their interactions. Therefore, as the

random forest proved to be the best fit for the model, the feature importance was calculated based on the method described in section 2.2.2 for this particular data mining technique, calculating the decrease in node impurity weighted by the probability of reaching that node.

The feature importance was thus calculated, and the results for the top 15 variables are displayed in a bar chart in Figure 4.5, where the highest scores represent the features that contributed the most towards the construction of the predictive model. It comes as no surprise that the variable that impacts the hexane concentration the most is the *IT_motor*, as it represents the current intensity of the main engine.

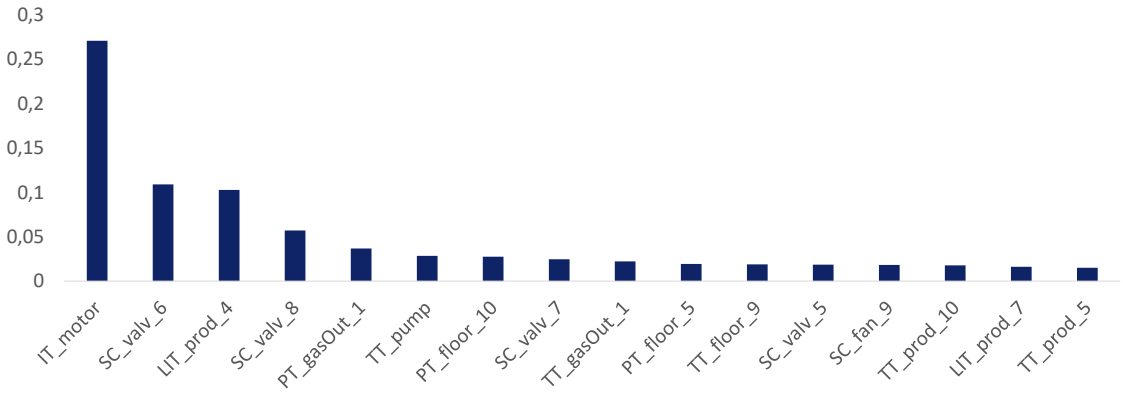

Figure 4.5: Bar chart representing feature importance in the random forest

Furthermore, it should be noted that the five most important features that resulted from this analysis, *IT_motor, SC_valv_6, LIT_prod_4, SC_valv_8*, and *PT_gasOut_1*, also presented a p-value under 0,05 in the stepwise regression model calculated earlier. Even though this approach fails to consider the interactions between the variables, the fact remains that these variables are considered the most relevant in both methods. However, it should be noted that there are no variables that present a very strong impact on the hexane concentration on their own. This outcome is explained by the thermodynamic equilibrium present in the DT/DC, which implies that all the variables interact and influence each other, as well as the solvent concentration present in the flour.

4.3.6 Operational Model

A second model was developed, aiming to define the optimal values of the input variables that guarantee the required hexane concentration. Using the group division defined in section 4.3.2, the maximum and minimum values of each group were established and are summarized in Table 4.7. The complete table is displayed in Table B.5 in Appendix B.

This model enables the operational supervisor to become aware of the optimal values of each parameter that lead to a hexane concentration under 500 ppm, as well as the values that lead to the other concentration ranges. The following conclusions were drawn for each variable from the analysis of Table 4.7:

Table 4.7: Minimum and maximum values of selected variables in each group

Variables	≤ 500		$]500,700]$		$]700,900]$		$]900,1100]$		> 1100	
	Min	Max	Min	Max	Min	Max	Min	Max	Min	Max
SC_fan_9	36	38	32	38	34	38	35	38	25	28
SC_fan_10	36	38	32	38	30	38	35	38	25	28
SC_valv_5	15	25	15	50	15	50	18	32	0	50
SC_valv_6	16	18	16	20	16	40	16	23	0	28
LIT_prod_4	37	48	37	52	3	49	37	52	2	51
LIT_prod_5	36	48	35	63	2	54	36	53	3	67
TT_gasOut_1	73	75	73	78	73	99	73	76	73	100
IT_motor	352	374	330	393	268	392	311	374	186	357
TT_prod_9	49	63	45	69	45	61	45	65	44	75

- *SC_fan_9* and *SC_fan_10*: a high hexane concentration is expected when the driving force of these blowers is below 36 Hz;
- *SC_valv_5*: a high hexane concentration is expected when the speed of the sluice valves is over 25 Hz;
- *SC_valv_6*: a high hexane concentration is expected when the speed of the sluice valves is over 18 Hz;
- *LIT_prod_4* and *LIT_prod_5*: a high hexane concentration is expected when the level of the flour surpasses 48 °gr ;
- *TT_gasOut_1*: a high hexane concentration is expected if the temperature of the gas that is leaving the equipment exceeds 75 °C;
- *TT_motor*: a high hexane concentration is expected when the main engine's current intensity goes below 352 A or surpasses 374 A;
- *TT_prod_9*: a high hexane concentration is expected if the temperature of the flour goes below 49 °C.

These findings enable the definition of the optimum setpoints for each parameter, as well as the determination of alerts and actions, that aligned with the operation and the defined thresholds, will support the production team in the process control and supervision.

4.4 Evaluation

The previous assessment steps analyzed the resulting model's accuracy and precision. This phase now aims to evaluate the degree to which the designed model meets the defined business goals. In fact, the total output of a data mining project is composed not only by the generated models, but also by the findings it triggers, namely all the important discoveries that meet the business objectives and lead to new questions, lines of approach, and potential side effects.

The first analytical model developed was based on the random forest algorithm, providing a real-time estimation of the final product's quality, with a prediction error of approximately 14%. In line with the projected business goals, the presented solution enables the operation supervisor to monitor the whole production process, with the support of real-time information concerning the extracted flours' hexane concentration, enabling the quality control of the end product. In fact, the developed model provides the client with entirely new information, since their current knowledge on the hexane concentration derives from weekly laboratory analyses, with results presented with an 8-hour delay.

Furthermore, this study proved that the underlining performance of machine learning models improves substantially with the amount of available data. As such, provided the client is able to collect more information, there is a potential to significantly diminish the current error of 14%, where the ultimate goal is to achieve a prediction error of approximately 0% for more accurate results. It should be noted that the resulting evidence is crucial to advert the client to the importance of performing consistent laboratory analyses, as one of the challenges inherent to digital transformations is the unequivocal resistance to change. Such resilience can be expected as manufacturers are conscientious of the risks inherent to digital transformations and are therefore reluctant to change the ongoing traditional methods that they trust and are familiar with. It is thus important to gather enough information that corroborates the fact that more data is required to increase the model's accuracy to enhance the real-time process supervision and product's quality control. The results from Table 4.6 prove this as the datasets with more observations are the ones with lower prediction errors.

Moreover, the second model enabled the definition of the optimum values for each operational parameter. This enables the supervisor to monitor each variable and understand whether their current values surpass the established limits. Furthermore, this operational model supports the definition of the optimum values for each setpoint that the operators should employ, rather than relying on the theoretical values based on the process's state of the art. As such, an opportunity arises to perform a fine-tuning of the process, since the supervisor is able to send adjustment orders to the operators based on the thresholds defined for the desired hexane concentration value.

In conclusion, the first two business goals outlined in section 3.5 were successfully accomplished. Together, the two analytical models provide an opportunity to calibrate the process in real-time. This approach naturally creates added value as the business currently struggles to obtain the legislated quality indicators of the produced flour. With the support of the developed models, along with its implementation in a technological solution, the production team will be able to monitor and control each parameter in due course, increasing the quality and safety indicators of the final product to the legislated values. The third business goal, regarding the referred implementation, will be accomplished in Chapter 5.

Chapter 5

Technology Meets Deployment

The deployment phase is the final stage in the CRISP-DM methodology. This chapter aims to expose the proposed deployment strategy, establishing how the developed models shall be implemented within the organization's systems and the benefits that arise from it.

5.1 Functional Architecture

A functional architecture discloses a high-level, simplified view of how the different systems and functions in a business operate and interact. The architecture for this project was designed based on the acquired knowledge of the client's landscape and is displayed in Figure 5.1. It includes two technological solutions, SAP Digital Manufacturing Cloud (DMC) and SAP Analytics Cloud (SAC). Powered by AI, these solutions augment human capabilities, allowing the development and implementation of analytical models to be paired with the automation of work orders, alert notifications, and readjustment orders.

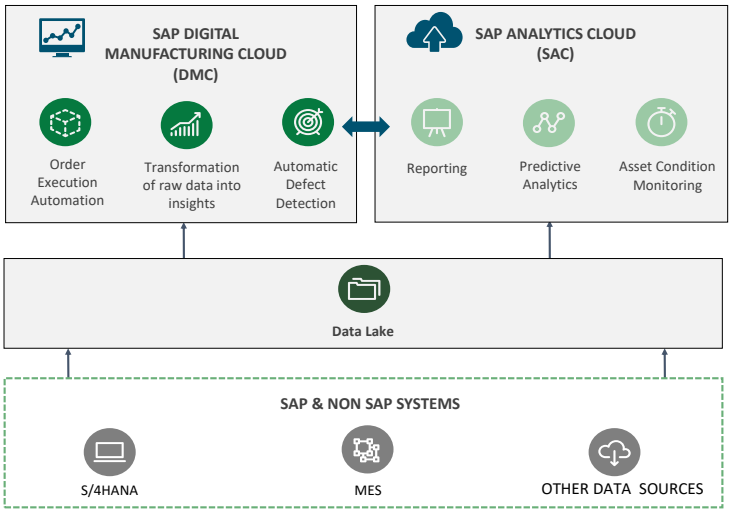

Figure 5.1: Functional Architecture

SAP systems such as S/4 HANA (an Enterprise Resource Planner (ERP)), or non-SAP systems such as Manufacturing Execution Systems (MES), and other data sources, can be used to feed a data lake, which is a repository of data stored in its natural/raw format. The DMC solution can retrieve the necessary data from the data lake and transform the information into valuable insights, leveraging on AI and ML capabilities, while enabling the automation of order executions and defect detection. SAC interacts with the DMC, reporting the respective data in visual dashboards, allowing the automation of shop floor operations. The solution also retrieves information from the data lake and other sources, leveraging its predictive analysis capabilities to analyze data and perform asset condition monitoring.

Due to the time constraints associated with the nature of this curricular dissertation, the implementation using SAC was considered the focus of this project.

SAC combines predictive analytics, business intelligence, and planning capabilities into a cloud-based data visualization tool. Its primary function is the creation of data reports, whose information originates from various sources, including real-time data captured from business activities and direct database/ERP readings, as well as data imports, namely regular excel spreadsheets.

For this dissertation project, the machine learning models were developed in an external environment, the results were exported onto an excel spreadsheet and then uploaded in SAC. However, in future deployment, the models will be embedded in the solution, automatically retrieving the required information from the database.

Furthermore, SAC exposes and highlights essential information through charts, tables, and other graphical components. As a result, the solution enhances asset condition monitoring, enabling the detection of anomalies and fluctuations that affect asset health and performance. It also ensures proactive, timely maintenance that prevents breaks and outages, while providing real-time visibility into asset health via automatic alerts and notifications.

5.2 Reporting

Taking into account the current as-is situation described in section 3.4.1, and in order to provide the different stakeholders with the predictive capabilities brought forth by machine learning, a set of dashboards were designed to address the roles and responsibilities of each of the following intervening personas in the process:

- *Production Director*- responsible for assessing the monthly production needs, as well as releasing requests for the laboratory analyses;
- *Production Supervisor* - answers to the production director and ensures the production orders are successfully carried out;
- *Operator* - answers to the director and supervisor, operating the machinery, defining/ monitoring the operating parameters, and collecting/ delivering flour samples to the laboratory.

A dashboard is a tool used to support information management and business intelligence, aiming to provide knowledge and understanding to all users, concerning the most critical analytics in the business, department, or specific process. The developed dashboards and the goals each one seeks to achieve are identified as follows:

1. **Primary Dashboard**: provides high-level, simplified information that enlightens the production director on the key points currently affecting the process;
2. **Laboratory Analyses Dashboard**: allows the production director and supervisor to access information regarding the past analyses on the flour's hexane concentration;
3. **Operational Parameters' Dashboard**: displays information regarding the current values of each operational parameter and their optimum values, enabling the operator and supervisor to monitor and control each parameter, while performing the necessary adjustments;
4. **Parameters Evolution Dashboard**: enables the operator and supervisor to visualize each parameter's evolution in time and its behaviour compared to its optimum boundaries.

Each dashboard was carefully designed, bearing in mind the respective users and the referred goals. As such, the most important key performance indicators (KPIs) were selected, along with the graphs that provide the best visual understanding of the displayed information. The devised dashboards allow each user to perform a quick scan and retrieve the most relevant information concerning the oilseed process, eliminating the need to sort through spreadsheets, emails, and the SCADA system.

5.2.1 Primary Dashboard

The primary dashboard, represented in Figure 5.2, exposes three KPIs, affected by date and time filters, that are present in all four dashboards. The first KPI, " Samples within the legislated limits

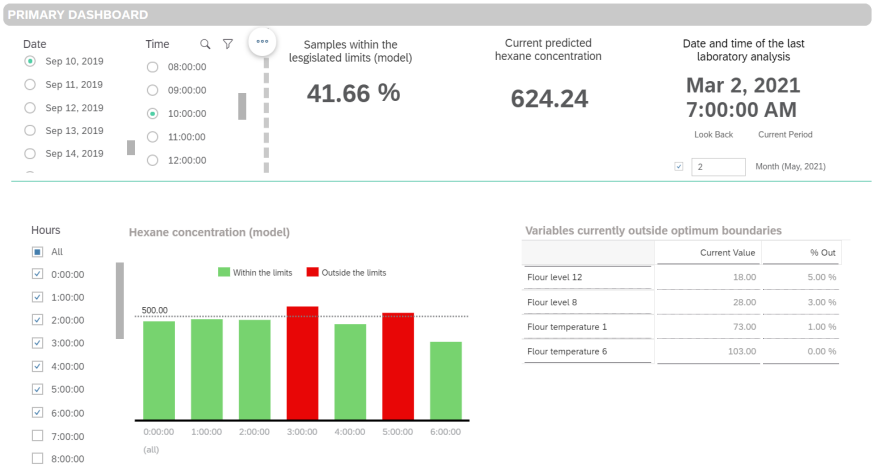

Figure 5.2: Primary Dashboard

(model)," calculates the number of samples that are below 500 ppm each day and displays the information as a percentage. This value is based on the predictions obtained from the model for each day individually. The following KPI, "Current predicted hexane concentration", shows the model's forecasted value of hexane concentration for the selected date and time. Finally, the last KPI shows the date and time of the last analyses carried out in the laboratory. The filter below allows the user to select the number of past months this information should include. In other words, if the user selects two months, all the dates from the analyses in the last two months are displayed.

Below the KPIs, a bar chart displays information on the predicted hexane concentration in the past hours. The green bars represent acceptable values, and the red depict concentrations over 500 ppm. The user can hover the mouse over the bars to visualize the actual values. He can also chose the desired number of past hours to be seen, using the filter displayed on the left. On the right side, the table shows the variables currently outside the calculated optimum boundaries, their current value, and the percentage by which they surpass the limits.

This dashboard informs the production director on whether the current production batch is foreseen to be within the legislated limits, which parameters currently require the most attention, and the resulting proportion of acceptable batches each day. As a result, the dashboard supports the assessment of production needs, illustrating how the past plan impacts the flour's quality.

5.2.2 Laboratory Analyses Dashboard

The laboratory analyses dashboard is used by both the production director and supervisor and is represented in Figure 5.3. It contains information regarding the past analyses on the flour's hexane concentration.

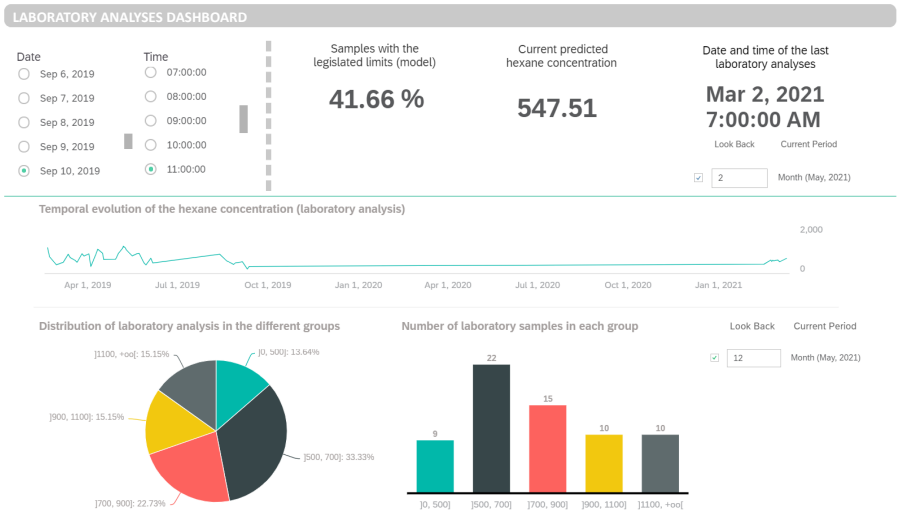

Figure 5.3: Laboratory Analysis Dashboard

This dashboard includes the same three KPIs as the previous one, as well as three new graphs. The first one exposes a time-series representation of all the analyses carried out in the laboratory.

On the bottom left corner, the pie chart shows how these samples are distributed in the different concentration groups mentioned in section 4.3.2. On the right side, the bar chart shows the number of laboratory samples existing in each group. Finally, the date filter screens the information of these graphs to the desired number of past months.

The production director and supervisor benefit from the displayed information as it provides knowledge on how the actual hexane concentration is evolving through time. The time-series graph illustrates whether the hexane concentration increases or decreases at different times, revealing long periods where no analyses were carried out. In addition, the pie and bar chart display whether the number/percentage of samples in high concentration groups has been decreasing or increasing in the past months. Moreover, the director is able to plan the need for the subsequent laboratory analyses based on the date of the previous ones.

5.2.3 Operational Parameters Dashboard

The third dashboard, represented in Figure 5.4, displays real-time information on the 42 operational parameters. In addition to the first pair of KPIs present in the previous dashboards, a new KPI is displayed on the top-right corner, indicating the number of parameters that are currently within the calculated limits for the selected date and time.

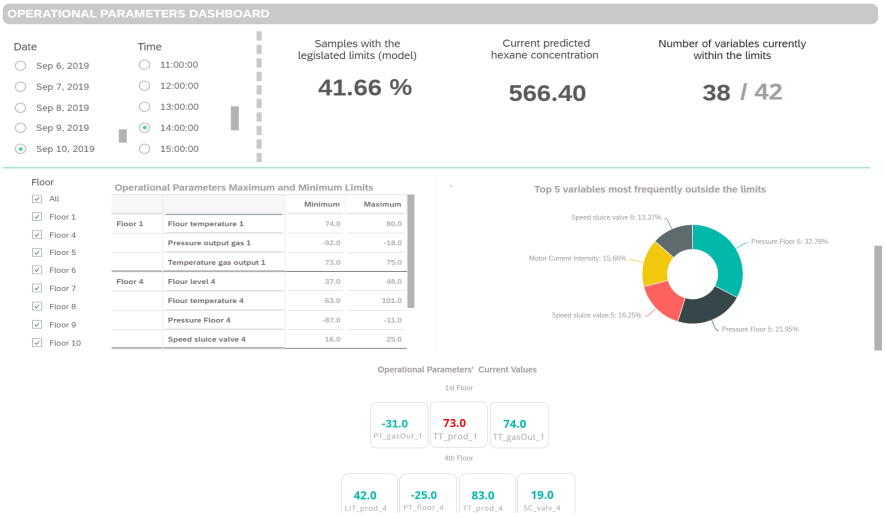

Figure 5.4: Operational Parameters Dashboard

The table below displays the maximum and minimum boundaries calculated for each variable, ordered by floor. This organization enables the operator or supervisor to monitor each floor sequentially, where the filter on the left side enables them to select the preferred floor. On the right side, the doughnut chart shows the top 5 variables most frequently outside the limits, enlightening the users to which variables require the most attention.

The current values of each parameter are displayed below, again organized by floor. When a parameter is within the calculated limits, it appears in green; otherwise, it appears in red.

In short, this dashboard enables the operator and supervisor to monitor and control the operational parameters. The new KPI immediately informs them whether there are any variables outside the optimum boundaries. They can then quickly pinpoint them as the red values, compare them to the optimum thresholds in the table, and adjust them in the equipment accordingly.

5.2.4 Parameters' Evolution Dashboard

Finally, the fourth dashboard is represented in Figure 5.5, illustrating temporal diagrams of each variable's evolution. It also includes the same three KPIs as the previous dashboard. Despite

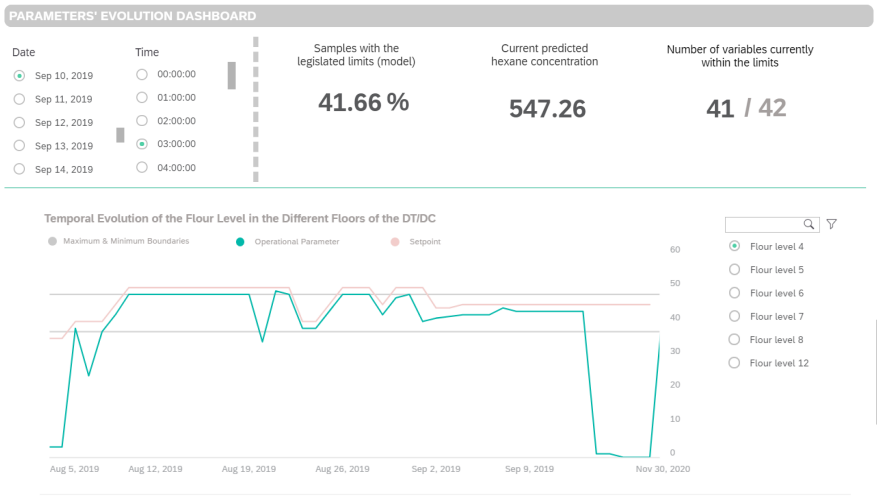

Figure 5.5: Parameters' Evolution Dashboard

what is seen in Figure 5.5, as the user scrolls down, all the parameters are duly represented. The variables are grouped by the type of object they monitor and a filter on the right side allows the user to select the desired floor. These graphs not only reveal each parameter's variation in time, but also present their optimal maximum and minimum boundaries, as well as the setpoint's variation. This enables the operator and supervisor to gain quick, visual insights on the patterns of behaviour of each variable and compare them to their ideal thresholds and target values..

These four dashboards are the resulting reports that arise from the developed machine learning models. They provide the different process stakeholders easy access to critical information that allows them to make faster and better decisions on how to manage shop floor operations, increasing efficiency and productivity. Furthermore, the SAC implementation brings the business' data onto the cloud, allowing users to keep track of the vital data anytime, anywhere. As such, it is clear that the use of artificial intelligence enhances the ability to monitor multiple KPIs in central dashboards, enabling the process fine-tuning in real-time. Ultimately, this wide-ranging branch of computer science leads to an inevitable expansion of human capacity, providing mankind a chance to extend the reach of human cognition and capability. The two worlds are thus compelled to merge and work together, as a means to improve one another mutually.

Chapter 6

Conclusion and Future Work

There are two main topics at the core of this curricular dissertation. To begin with, the agrofood industry is exposed to rigorous external enforcement regarding quality control procedures. On top of that, the high degree of dependence on human activity to control the production process incites a high variability in solvent extraction operations. Faced with an unstable desolventization process, the company under study recognized that the exceeded hexane concentration in the extracted flours causes a significant impact on the final product's quality and safety indicators.

The developed project proposes an intervention and optimization of the desolventization process, aiming to reduce the hexane concentration of the extracted flours and enable the process control and calibration in real-time. The primary objective was to comply with the legislated quality and safety standards.

During the project's development, thorough analyses and clarifications regarding the process and the results occurred in work sessions, in the presence of the company's production team. The purpose of these events was to incorporate the know-how and expertise of the production team specialized in the solvent extraction process, in order to ensure the definition of suitable goals and proposed solutions regarding the organization's strategy. The participation of the production team played a major role in understanding the business and clarifying important issues and assumptions that arose during the course of the project.

6.1 Main Conclusions

The first phase of the project involved understanding the oilseed business, as well as the desolventization process, the equipment used, and the relevant variables. As such, the process was acknowledged as a thermodynamic system, where the temperature, flour levels, steam, and pressure are all in equilibrium, meaning that all the variables interact with each other and together play an important role in the final hexane concentration.

Next, the operational and laboratory data was thoroughly analyzed, adverting to the existence of data quality issues. To begin with, the performed statistical analysis uncovered the need to treat and prepare the data for the preparation phase, due to the high number of outliers that resulted

from production breaks at the factory. Regarding the laboratory measurements, the method used to collect this data proved to be inefficient, as these observations result from manual sample collections and subsequent laboratory analyses, whose results arrive with a significant delay, obstructing the real-time process tuning. Furthermore, these analyses take place between wide time periods, whereas the operational data is automatically recorded every 5 seconds. As a result, the number of laboratory observations is quite small when compared to the operational data. Furthermore, the statistical analysis proved the lack of significant variation in the DT/DC operational parameters in the course of one hour, which in agreement with the client, lead to the decision of aggregating the data in an hourly interval, based on the median values of each period.

After the data aggregation and outlier treatment, three datasets were constructed, where one contained only original laboratory values, and the other two estimated the missing data points. One of the estimates took into account the assumption, confirmed by the client, that there are no signs of variation in the hexane concentration in the course of one hour. Therefore, the hexane concentration one hour before and after each observation was assumed to remain the same, resulting in a dataset with thrice as many observations as the original one. On the other hand, the other dataset was computed using the KNN algorithm, relying on the values of the closest neighbours to predict the missing values. However, given the small nature of the original data, the end results proved to be misleading, since the training data for the future models used previously predicted data points, rather than original observations. Nevertheless, the KNN dataset still succeeded in demonstrating that the increasing training data lead to decreasing prediction errors.

Once the models were constructed and evaluated, the random forest proved to be the best technique for this project, with the highest prediction accuracy. Meanwhile, the results from all the techniques for the different datasets proved that the model errors decreased with increasing data, meaning that more data is required for the model to yield perfect results, i.e., an error approximately equal to 0%.

After discussing the final model's results with the client, it was found that, although not perfect, the model's performance was excellent given the current variability of the desolventization process. The analytical models were able not only to provide a real-time prediction of the hexane concentration with a low prediction error, but also determine the optimum thresholds for each operational parameter, providing entirely new and valuable information for the client.

Furthermore, regarding feature importance, it became clear that, even though it is possible to determine the features that contributed the most towards the construction of the model, there are no variables that strongly impact the hexane concentration on their own. This results from the fact that the DT/DC is a thermodynamic system in equilibrium, where all the variables interact and influence one another.

The final stage of the project addressed the implementation of the analytical models in interactive dashboards for the different personas engaging in the process. A functional architecture was designed to embed the developed models in a technological solution and generate hands-on interfaces that enable the production director, supervisor, and operator, to interact with the results obtained from the models. The displayed KPIs and graphs enable the user to evaluate the

process's condition and depict whether it is operating as expected, or if certain adjustments are required. Information regarding the current vs. optimal state of operational parameters, past laboratory analyses results, and real-time hexane concentration predictions provide a ground-breaking opportunity for the client to monitor and control the parameters, while performing the process fine-tuning in real time, which they were unable to do before.

In the end, the project succeeded in meeting the quality and safety standards, while radically reshaping the way the production team analyzes, evaluates, and interacts with the process. The developed analytical models enabled the optimization of the desolventizing process and shop floor operations, by increasing the efficiency and effectiveness of parameter monitoring and control, derivative of the real-time hexane concentration predictions. The integration of the proposed solution yields benefits that would otherwise not be possible, such as the anticipation of production issues and real-time process calibration.

The core of this solution is built on the combination of multiple emerging Industry 4.0 technologies, whose results revealed an undeniable impact on the organization, revolutionizing the entire process culture and work routine. A significant change is expected in the way workers, processes, and machines interact to generate efficiency and attract new sources of innovation, to support informed decision making, and increase productivity and agility in problem-solving. Furthermore, it should be noted the solution is scalable to other processes within the business and other company plants, empowering the business' sustainable growth and ability to adapt to incoming adversities.

6.2 Future Work

While analyzing this project's main findings, several opportunities arise for further process optimization and possible future work. One of the shortcomings identified in the course of this dissertation was the reduced amount of available data to feed the machine learning model. As such, it is important that the client is able to perform laboratory analyses on the hexane concentration more frequently in the future. This way, more data can be used to train the model, leading to lower and lower prediction errors, where the aim is to achieve an error of approximately 0% for more accurate results. This continuous record will allow the development of increasingly reliable forecast models that estimate the trend of each variable in real-time, according to the historical records.

Furthermore, this project enabled the definition of optimal values for the model's input variables, according to the hexane concentration set in the legislated quality and safety standards. As such, the optimization of these values could potentially increase energy efficiency. Given that the desolventization process is responsible for over half of the whole plant's energy expenditure, there is a great opportunity to significantly reduce overall costs. Therefore, a thorough study should be carried out in the future in order to perceive which variables directly impact energy expenditure and how these can be optimized.

Meanwhile, the developed models, in conjunction with the designed dashboards, will empower the definition of actions that enable the process optimization and refinement of shop floor operations. These will be based on the optimum thresholds defined for each parameter according to the different ranges of hexane concentration. Once the predictive models are implemented and optimized, a new analytical model can be developed to learn the recommendations and automatically inform the operators on the actions to be taken.

In order to further optimize the automation of the procedural tasks, the functional architecture previously displayed in Figure 5.1 should be implemented. This project focused solely on the implementation of the analytical models in the SAC solution due to the timeframe inherent to the nature of the curricular dissertation. However, the DMC solution, paired with the data lake and SAC, presents a great potential to optimize and automate the shop floor operations further.

The DMC is a tool that nurtures global visibility regarding the entire company's activities; it is designed to establish the link between production and business operations in supply chain management, increasing visibility between top floor management activities and shop floor equipment. The solution enables the collection and analysis of data from the production process directly from the factory floor. The data lake is a centralized repository that extracts all the data from the SCADA system and consolidates the information all in one place, generating a single source of truth (SSOT). This concept refers to the management of data in one singular place, ensuring everyone in the organization bases their business decisions on the same data. In parallel, the data lake enables the model procedures to be carried out and then feeds the results to the DMC and SAC solutions.

The DMC operates in two distinct areas, Manufacturing Execution and Manufacturing Insights, which support near real-time decision making. The DMC Execution has the ability to organize and control the factory floor, implementing real-time analytical solutions that provide better performance of production operations across distinct levels of the company hierarchy. It enables the use of intuitive interfaces for operators, such as 3D instruction manuals, maintenance documents, and alerts that suggest actions to be taken in the process. It also allows the automation of data collection and definition of process parameters based on analytical models. Moreover, the DMC Execution is flexible to short-term changes, as it schedules and issues production orders, considering labor restrictions, resources, and maintenance plans. It also provides a shop floor designer to model and implement production processes and automate sequences on the shop floor, based on top management business rules and defined parameters.

The DMC Insights, on the other hand, provides greater visibility into the performance of the production process in real-time and digital format, supporting informed decision-making based on data from the various operating systems. It provides a perception of the performance and productivity across the various company levels (different regions, factories, resources) and empowers informed decisions combining data from shop floor systems with execution systems, as well as information from other sources (e.g., ERP). Its custom design of interactive dashboards allows operators to analyze the process in real-time through relevant indicators such as equipment status, production evolution, and alarms. Furthermore, the DMC Insights enables the analysis of the root

causes and failure modes of the process, allowing the user to trace upcoming problems across the various levels of the organization, and providing the opportunity to predict quality defects in almost real-time with the support of analytical models.

A projected user journey that entails the use of SAC and DMC is presented in Figure 6.1, where the distinct icons indicate which solutions support each step. The journey begins with the definition of the operational parameters. With the support of analytical models, the user is able to analyze the optimum parameters and automate the operationalization. For this purpose, the operator has access to textual and 3D work instructions that provide visual information at specific points in the manufacturing process, providing assistance with the necessary tools.

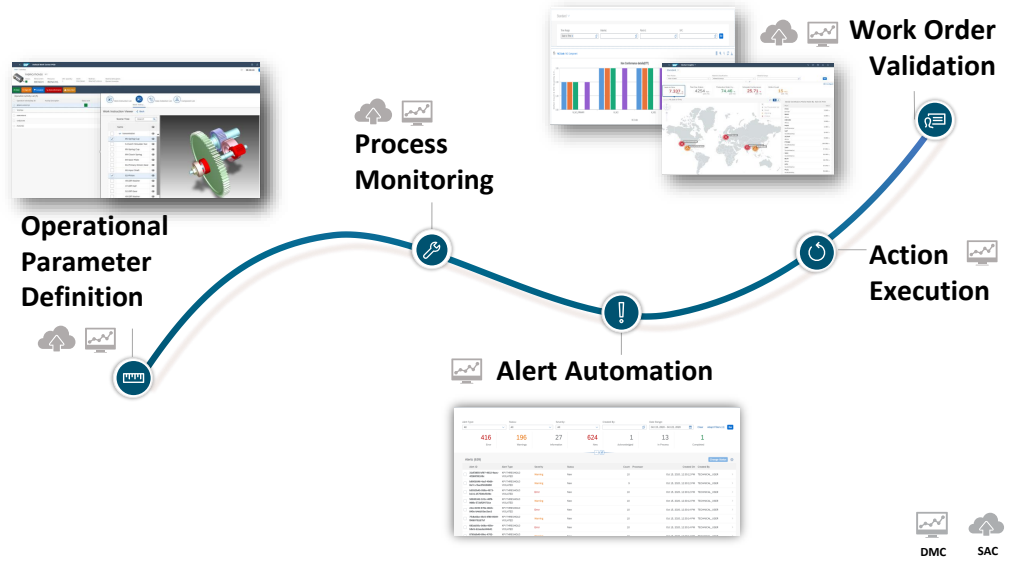

Figure 6.1: User Journey

Next, the process monitoring is performed with the support of alerts disputed by the solution via business rules. The person in charge has the opportunity to analyze the current issues inherent to the process and create work or process orders to address them, where instruction manuals and maintenance documents are available to guide the handling of the equipment.

Subsequently, a set of alerts can be automated based on thresholds and business rules, notifying the user of the recommended set of actions to be performed. The notified users can be selected by the operator or supervisor based on their responsibility to take action.

The supervisor then analyzes and releases these actions to the operators, if deemed necessary. Once the actions have been released by the person in charge, the operator receives the work orders and executes them, confirming the order completion in the end.

Finally, the work order validation is performed with the support of the reporting layer that embeds the analytical models. The information displayed on these dashboards ranges from high-level, top management reports, down to the operating/equipment status. Furthermore, the production director has access to the results in the different company plants and can therefore gain real-time visibility and monitor the performance of the production process from a regional level through

geographic mapping. This enables him to compare the performance between plants within and across different regions.

In the end, the projected goals were successfully achieved, as the approach proved to adequately optimize the solvent extraction process and improve the final product's quality indexes. Given the underlying potential scalability of the tool, the future replication of the solution is expected, as the remaining industrial processes can also be digitally transformed.

Nevertheless, there are numerous challenges imposed by digital transformations in a company. An open-minded and unique awareness is required to enable the shift from the conventional, old-school procedures towards connected and intelligent machines, processes, and overall environment. Some resistance to change can be expected as this new culture revolutionizes the traditional working routines that date back many years. However, the deployment of these intelligent systems testifies to the potential of artificial intelligence in augmenting and extending human cognition and abilities, as it provides information to the operators that could not have been achieved otherwise. As such, the sole focus on the optimization of specific shop floor operations was crucial to advert the client towards the potential of Industry 4.0 technologies and the possibility of implementing gradual changes, before escalating and expanding to the remaining industrial processes and manufacturing plants.

Bibliography

Aba, E. K. and Hayden, M. A. (2013). Understanding variation and its relationship to quality.

Anand, R., Kanagachidambaresan, G. R., Balasubramanian, E., and Mahimat, V. (2020). *Internet of Things for Industry 4.0 : Design, Challenges and Solutions*. Apress.

Asbroeck, B. V., Debussche, J., and César, J. (2019). Big data & issues & opportunities: Data ownership. *Bird & Bird*.

ATKearney (2018). The data value chain. Technical report, GSMA.

Balamurugan, E., R.Flaih, L., Sangeetha, K., Yuvaraj, D., Jayanthiladevi, A., and Kumar, T. S. (2019). Use case of artificial intelligence in machine learning manufacturing 4.0. *International Conference on Computational Intelligence and Knowledge Economy*.

Burke, B. (October 19, 2020). Gartner top strategic technology trends for 2021.

Chen, V. C. P., Kim, S. B., Oztekin, A., and Sundaramoorthi, D. (2018). Data mining and analytics. *Annals of Operations Research*.

Choudhury, A., Behl, A., Sheorey, P. A., and Pal, A. (2021). Digital supply chain to unlock new agility: a tism approach. *Benchmarking: An International Journal*.

Columbus, L. (2018). 10 ways machine learning is revolutionizing supply chain management. *Forbes*.

Columbus, L. (2020). Roundup of machine learning forecasts and market estimates, 2020. *Forbes*.

Cotteleer, M. and Sniderman, B. (2017). Forces of change: Industry 4.0. *Deloitte*.

Deloitte (2020). Available in: `https://www2.deloitte.com/pt/pt.html`. Accessed: 2021-03-02.

Deloitte (2021). Tech trends 2021. Annual report.

do Campo, V. (2020). Agroalimentar | agroindústria | grande entrevista. Accessed: 2021-03-30.

Dogan, A. and Birant, D. (2021). Machine learning and data mining in manufacturing. *Expert Systems with Applications*.

D'Souza, S., V., P. K., and S, B. (2020). Feature selection and modeling using statistical and machine learning methods. In *2020 IEEE International Conference on Distributed Computing, VLSI, Electrical Circuits and Robotics (DISCOVER)*.

ENEI (2014). Diagnóstico de apoio às jornadas de reflexão estratégica. Accessed: 2021-03-29.

Erol, S. and Sihn, W. (2017). Intelligent production planning and control in the cloud – towards a scalable software architecture. *Procedia CIRP*. 10th CIRP Conference on Intelligent Computation in Manufacturing Engineering - CIRP ICME '16. [Edited by: Roberto Teti, Manager Editor: Doriana M. D'Addona].

FIPA (2020). Evolução das exportações. Accessed: 2021-03-29.

Fui-Hoon Nah, F. and Siau, K. (2020). Covid-19 pandemic - role of technology in transforming business to the new normal. *The Institution of Engineering and Technology*.

Gilchrist, A. (2016). *Industry 4.0: Industrial Internet of Things*. Apress.

Gorunescu, F. (2011). *Data Mining - Concepts, Models and Techniques*. Springer.

Haq, A. U., Zhang, D., Peng, H., and Rahman, S. U. (2019). Combining multiple feature-ranking techniques and clustering of variables for feature selection. *IEEE Access*.

He, S.-G., He, Z., Wang, G., and Li, L. (2009). Quality improvement using data mining in manufacturing processes.

INE (2020). Estatísticas agrícolas. Accessed: 2021-03-30.

Karmakar, A., Dey, N., Baral, T., Chowdhury, M., and Rehan, M. (2019). Industrial internet of things: A review. In *2019 International Conference on Opto-Electronics and Applied Optics (Optronix)*. IEEE.

Kemper, T. G. (2020). Meal desolventizing, toasting, drying and cooling. Accessed: 2021-03-30.

Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., Wirth, R., Chapman, P., and Cinton, J. (2000). Crisp-dm 1.0 step-by-step data mining guide. Technical report, CRISP-DM Consortium, August 2000.

Kerner, S. M. (September 10, 2019). Top 8 cloud data warehouses. *Datamation*.

Kuhn, M. and Johnson, K. (2019). *Feature Engineering and Selection: A Practical Approach for Predictive Models*. Taylor & Francis Group.

Kusiak, A. and Salustri, F. A. (2007). Computational intelligence in product design engineering: Review and trends. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*.

Li, J., Tao, F., Cheng, Y., and Zhao, L. (2015). Big data in product lifecycle management. *The International Journal of Advanced Manufacturing Technology*.

Lu, S., Xu, C., Zhong, R. Y., and Wang, L. (2017). A rfid-enabled positioning system in automated guided vehicle for smart factories. *Journal of Manufacturing Systems*.

Magomadov, V. S. (2020). The industrial internet of things as one of the main drivers of industry 4.0. *IOP Conference Series Materials Science and Engineering*.

Mahmood, Z. (2019). *The Internet of Things in the Industrial Sector*. Springer.

Mohajan, H. K. (2019). The first industrial revolution: Creation of a new global human era. *Journal of Social Sciences and Humanities*.

Mohajan, H. K. (2020). The second industrial revolution has brought modern social and economic developments. *Journal of Social Sciences and Humanities*.

Mussomeli, A., Gish, D., and Laaper, M. (2016). The rise of the digital supply network industry 4.0 enables the digital transformation of supply chains. *Deloitte*.

Nicoletti, B. (2020). *Procurement 4.0 and the Fourth Industrial Revolution: The Opportunities and Challenges of a Digital World*. Palgrave Macmillan.

Nofal, H. (2019). The unspoken truth: The role of cybersecurity in breaking the digital transformation deadlock. GBM 8th Annual Security Survey.

North, K. and Varvakis, G. (2016). *Competitive Strategies for Small and Medium Enterprises: Increasing Crisis Resilience, Agility and Innovation in Turbulent Times*. Springer.

Nwokeji, J. C., Stachel, R., Holmes, T., Aqlan, F., Udenze, E. C., and Orji, R. (2019). Panel: Addressing the shortage of big data skills with inter-disciplinary big data curriculum. In *2019 IEEE Frontiers in Education Conference (FIE)*.

Ogorodnyk, O., Lyngstad, O., Larsen, M., and Martinsen, K. (2021). *EcoDesign and Sustainability I: Sustainable Production, Life Cycle Engineering and Management*, chapter Prediction of Width and Thickness of Injection Molded Parts Using Machine Learning Methods. Springer.

Onel, M., Beykal, B., Ferguson, K., Chiu, W. A., McDonald, T. J., Zhou, L., House, J. S., Wright, F. A., Sheen, D. A., and Rusyn, I. (2019). Grouping of complex substances using analytical chemistry data: A framework for quantitative evaluation and visualization. *Plos One*.

Ozdogru, U. (2020). Chapter 3 - impact of exponential technologies on global supply chain management. In Pagano, A. M. and Liotine, M., editors, *Technology in Supply Chain Management and Logistics*. Elsevier.

Parente, M., Figueira, G., Amorim, P., and Marques, A. (2020). Production scheduling in the context of industry 4.0: review and trends. *International Journal of Production Research*.

Pessoa, M. V. P. and Becker, J. M. J. (2020). Smart design engineering: a literature review of the impact of the 4th industrial revolution on product design and development. *Research in Engineering Design*.

Ponemon Institute (2018). Bridging the digital transformation divide : Leaders must balance risk & growth. IBM.

Ribeiro, V., Rocha, A., Peixoto, R., Portela, F., and Santos, M. F. (2017). Importance of statistics for data mining and data science. In *2017 5th International Conference on Future Internet of Things and Cloud Workshops (FiCloudW)*, pages 156–163.

Rifkin, J. (2011). *The Third Industrial Revolution: How Lateral Power Is Transforming Energy, the Economy, and the World*. Macmillian.

Ronaghan, S. (2018). The mathematics of decision trees, random forest and feature importance in scikit-learn and spark. *Towards Data Science*.

Rostami, H., Blue, J., and Yugma, C. (2016). Equipment condition diagnosis and fault fingerprint extraction in semiconductor manufacturing. In *2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA)*.

SAP (2020). Turning data into business value with sap business technology platform. Technical report, SAP.

Scholkmann, A. B. (2021). *Resistance to (Digital) Change*, pages 219–236. Springer International Publishing, Cham.

Serror, M., Hack, S., Henze, M., Schuba, M., and Wehrle, K. (2020). Challenges and opportunities in securing the industrial internet of things. *IEEE Transactions on Industrial Informatics*.

SISAB (2017). Setor agro-alimentar. Accessed: 2021-03-29.

Sniderman, B., Mahto, M., and Cotteleer, M. (2016). Industry 4.0 and manufacturing ecosystems: Exploring the world of connected enterprises. *Deloitte*.

Statista (2021). Avalibale in: `https://www.statista.com/statistics/250479/big-four-accounting-firms-global-revenue/#statisticContainer`. Accessed: 2021-03-02.

Tao, F., Qi, Q., Liu, A., and Kusiak, A. (2018). Data-driven smart manufacturing. *Journal of Manufacturing Systems*.

Uzzaman, A. (December 15, 2020). Top business and technology trends in 2021. *Inc*.

Vijayaraghavan, V. and Leevinson, J. R. (2019). *The Internet of Things in the Industrial Sector*, chapter Internet of Things Applications and Use Cases in the Era of Industry 4.0. Springer.

Walsh, T., Umbenhauer, B., Mussomeli, A., and Clark, J. (2018). Digital supply networks in industrial products manufacturing. *Deloitte*.

Wang, Y., Ma, H.-S., Yang, J.-H., and Wang, K.-S. (2017). Industry 4.0: a way from mass customization to mass personalization production. *Advances in Manufacturing*.

Williams, O. C. and Olajide, F. (2020). A technological approach towards the measurement of enterprise agility. *15th Iberian Conference on Information Systems and Technologies (CISTI)*.

Xu, L. D., Xu, E. L., and Li, L. (2018). Industry 4.0: state of the art and future trends. *International Journal of Production Research*.

Zhang, Z., He, X., and Kusiak, A. (2015). Data-driven minimization of pump operating and maintenance cost. *Engineering Applications of Artificial Intelligence*.

Zidek, K., Maxim, V., and Jan Pitel, A. H. (2016). Embedded vision equipment of industrial robot for inline detection of product errors by clustering–classification algorithms. *International Journal of Advanced Robotic Systems*.

Zin, J. and Vogel-Heuser, B. (2019). A qualitative study of industry 4.0 use cases and their implementation in electronics manufacturing. *IEEE*.

Appendix A

Agrofood Sector, Oilseed Process and Business Understanding

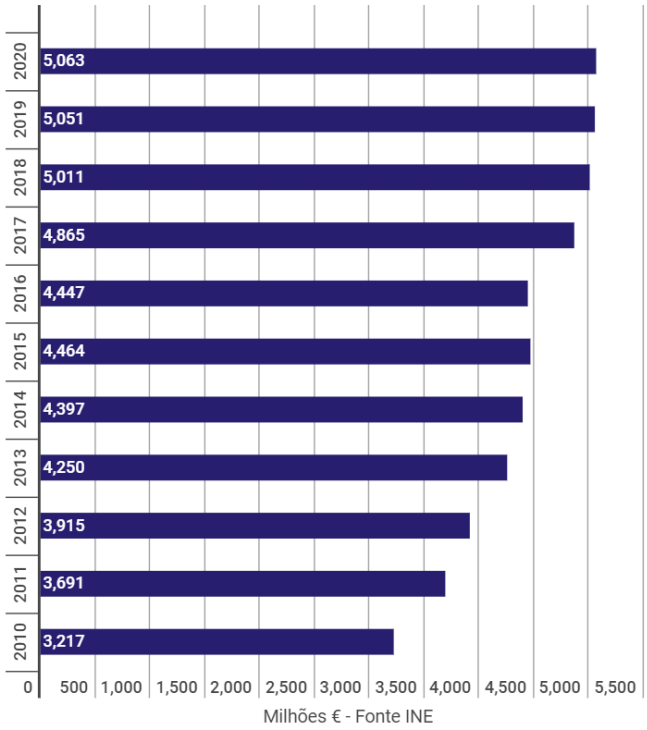

Figure A.1: Agrofood Industry Exports Evolution 2010-2020. Source: FIPA (2020)

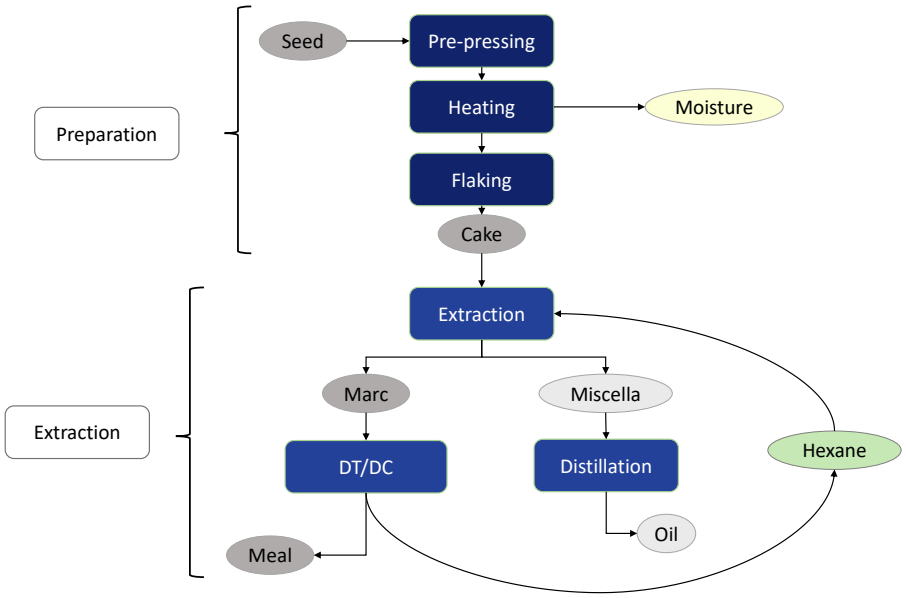

Figure A.2: Oilseed Processing, adapted from: Kemper (2020)

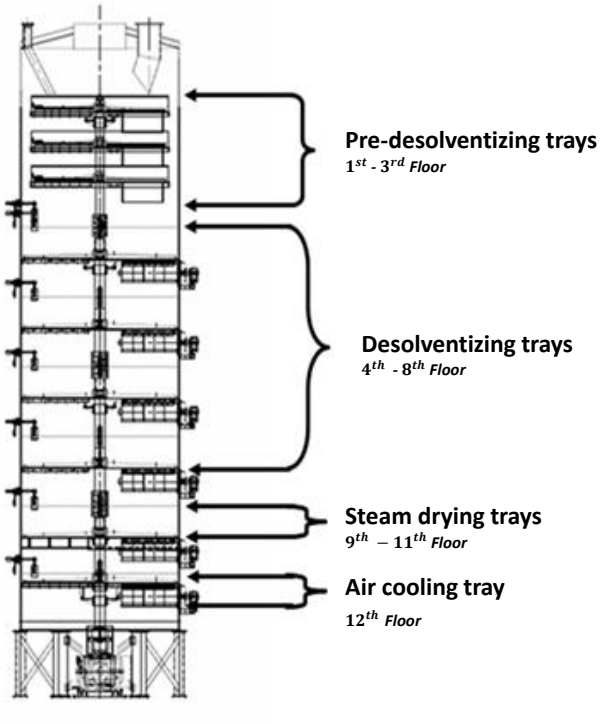

Figure A.3: DT/DC Schematic Representation. Source: Kemper (2020)

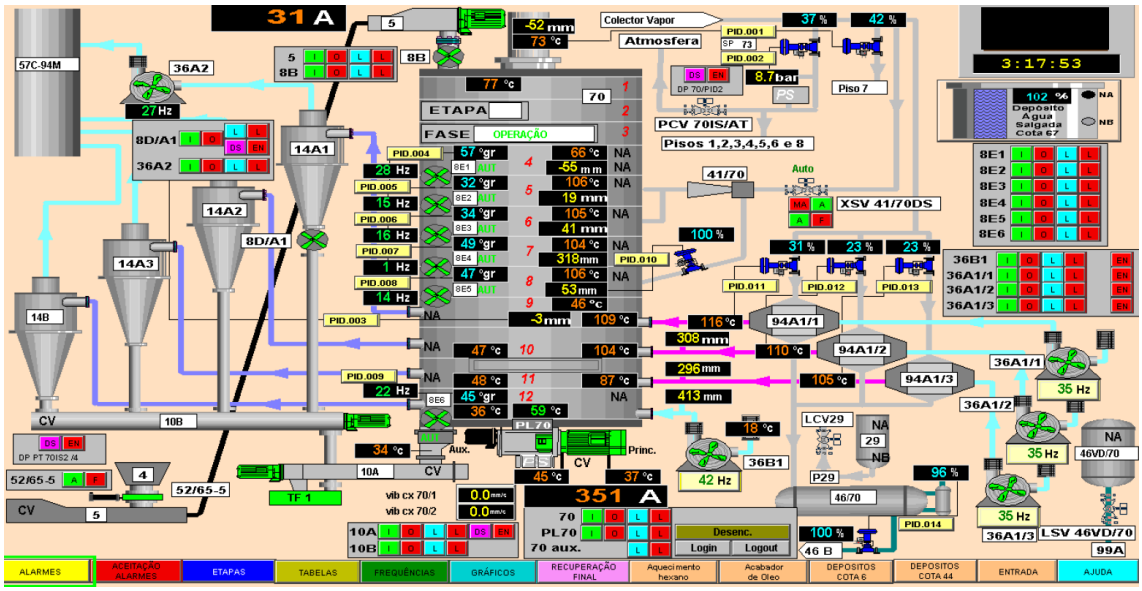

Figure A.4: Client's DT/DC Supervisory System, SCADA

Appendix B

Model Development Documentation

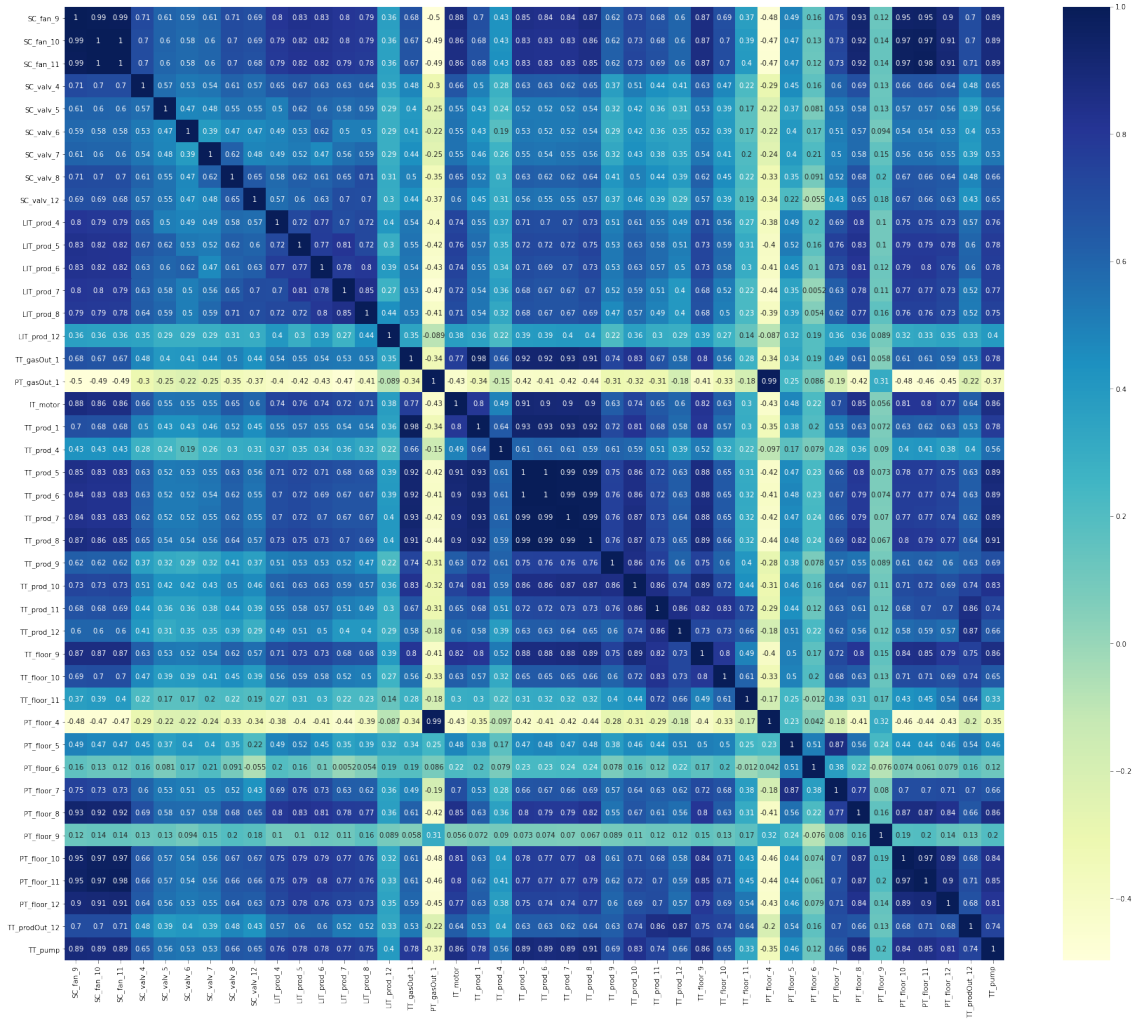

Figure B.1: Correlation Matrix Before Outlier Treatment

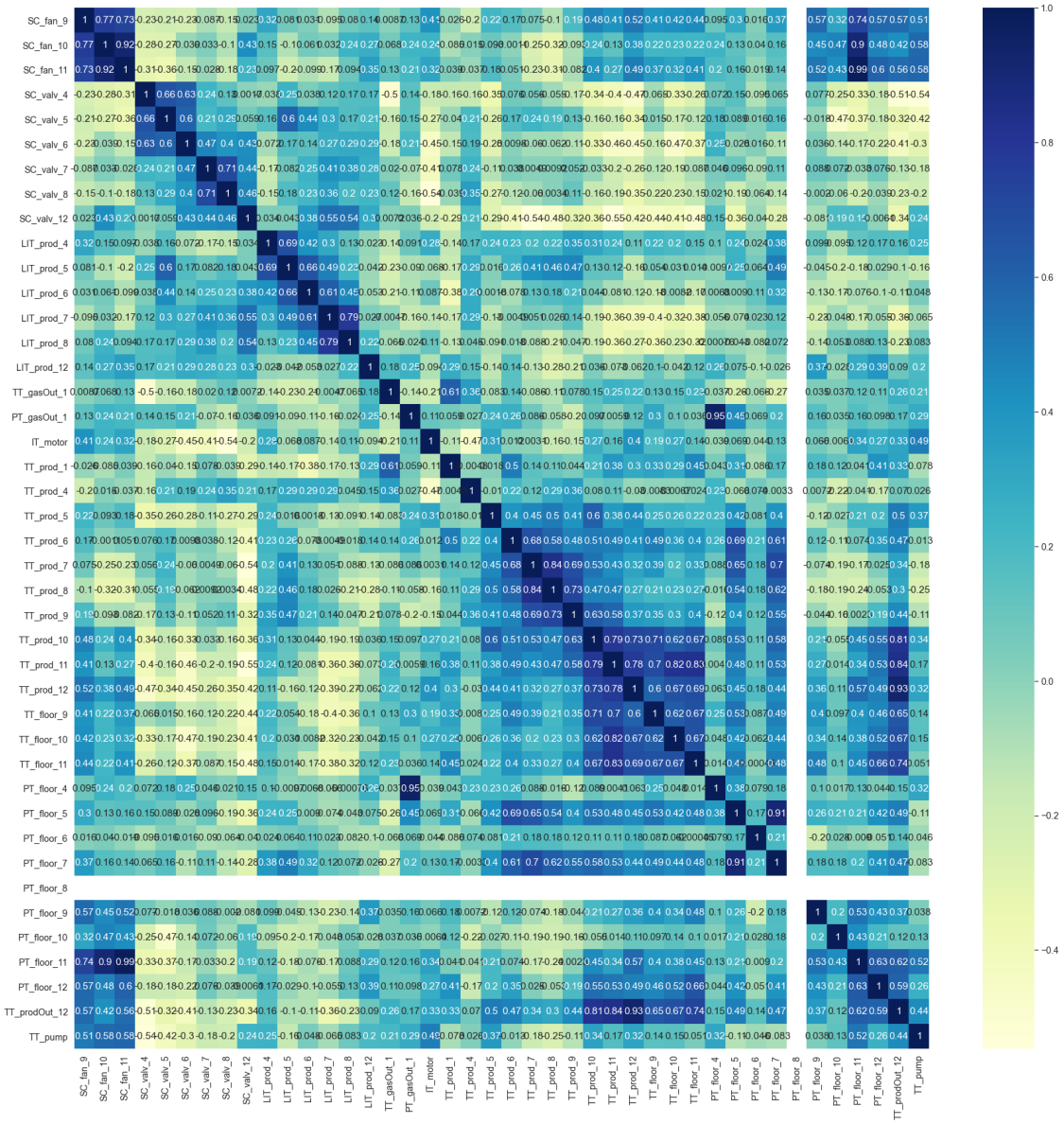

Figure B.2: Correlation Matrix of Dataset A

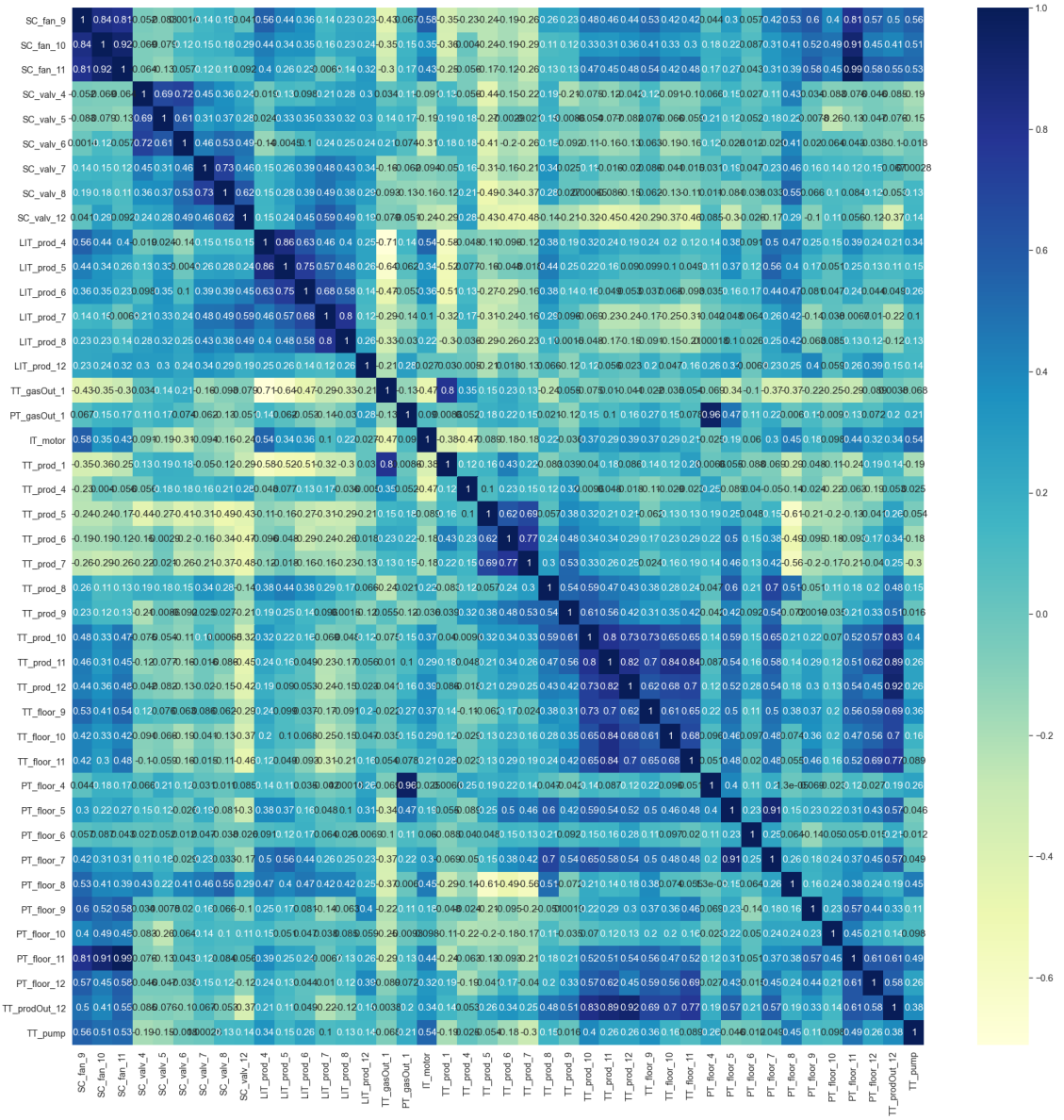

Figure B.3: Correlation Matrix of Dataset B

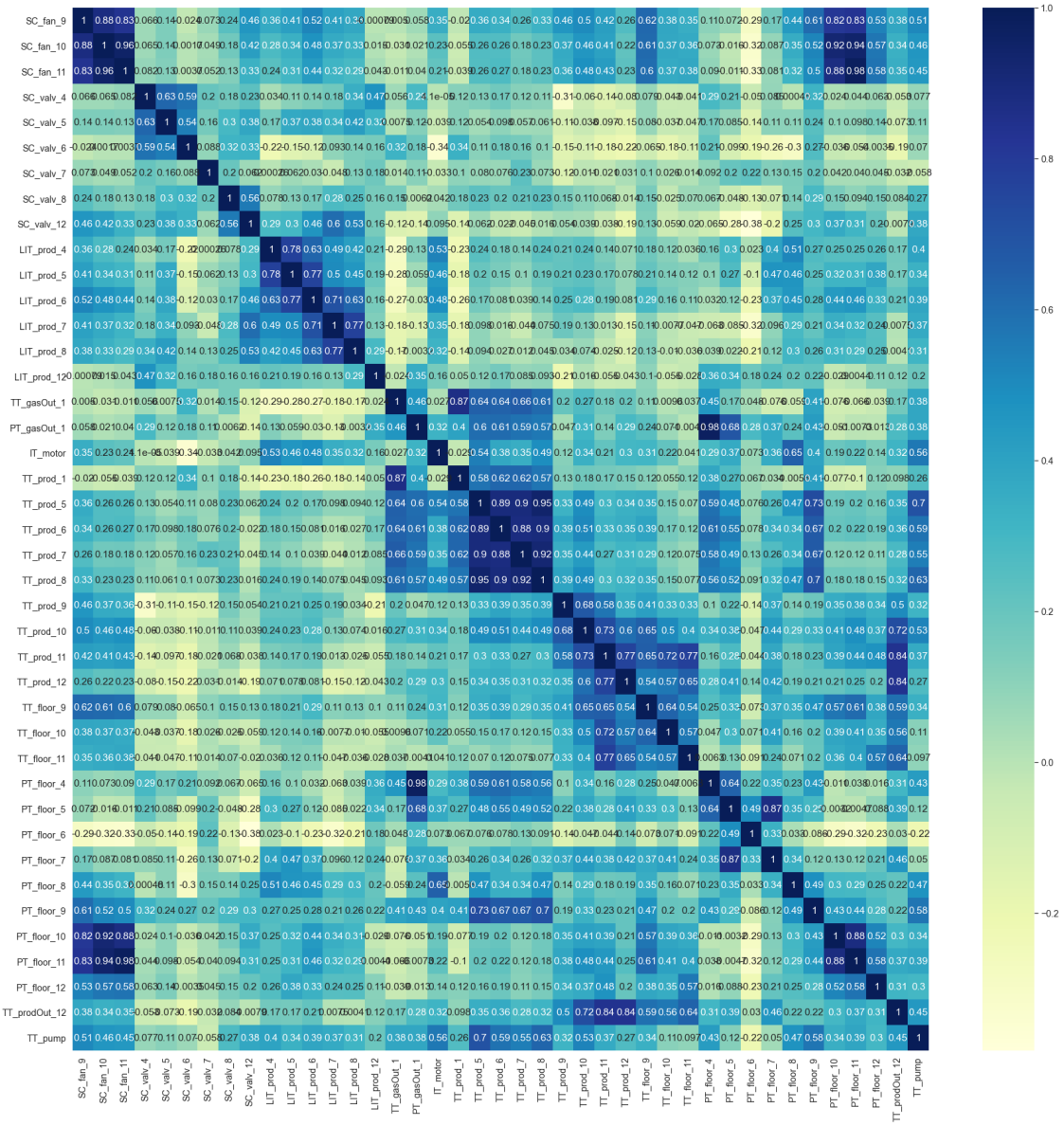

Figure B.4: Correlation Matrix of Dataset C

Table B.1: Variables selected by the stepwise regression forward Algorithm

Dataset	Included Variable	P-Value
Dataset A	IT_motor	0,000
	PT_floor_8	0,000
	LIT_prod_4	0,001
	SC_valv_6	0,002
	TT_floor_10	0,002
	TT_gasOut_1	0,008
	SC_valv_12	0,023
Dataset B	IT_motor	0,000
	SC_valv_7	0,000
	PT_floor_12	0,002
	PT_floor_7	0,002
	TT_prod_5	0,003
	TT_pump	0,003
	PT_gasOut_1	0,004
	LIT_prod_4	0,006
	TT_prod_7	0,008
	PT_floor_5	0,013
	LIT_prod_6	0,014
	SC_fan_9	0,017
	TT_prod_6	0,019
	PT_floor_11	0,024
	TT_prod_1	0,043
	SC_valv_8	0,046
Dataset C	IT_motor	0,000
	SC_valv_12	0,000
	PT_floor_6	0,000
	TT_prodOut_12	0,000
	PT_floor_8	0,000
	TT_pump	0,000
	TT_floor_11	0,000
	TT_prod_11	0,000
	TT_prod_10	0,000
	TT_prod_6	0,000
	PT_floor_7	0,000
	TT_floor_9	0,000
	PT_floor_10	0,000
	SC_valv_6	0,000
	TT_gasOut_1	0,000
	PT_gasOut_1	0,000
	PT_floor_4	0,000
	TT_floor_10	0,001
	PT_floor_5	0,007
	SC_valv_7	0,012

Table B.2: Variables selected by the stepwise regression backward Algorithm

Dataset	Included Variable	P-Value
Dataset A	LIT_prod_4	0,000
	SC_valv_12	0,000
	IT_motor	0,002
	SC_valv_6	0,002
	TT_floor_10	0,002
	SC_fan_10	0,004
	TT_prod_8	0,004
	SC_fan_11	0,006
	TT_gasOut_1	0,006
	TT_prod_5	0,049
Dataset B	SC_fan_10	0,000
	SC_fan_11	0,000
	TT_prod_8	0,000
	TT_prod_10	0,000
	SC_valv_12	0,000
	LIT_prod_4	0,000
	TT_gasOut_1	0,000
	IT_motor	0,000
	PT_floor_8	0,001
	TT_floor_10	0,002
	TT_floor_9	0,008
	TT_prod_5	0,011
	TT_prod_7	0,010
	PT_floor_10	0,028
	SC_valv_7	0,038
Dataset C	IT_motor	0,000
	SC_valv_6	0,000
	SC_valv_8	0,042
	SC_valv_12	0,000
	LIT_prod_4	0,011
	LIT_prod_5	0,065
	TT_prodOut_12	0,000
	TT_prod_11	0,000
	TT_prod_10	0,000
	TT_prod_5	0,000
	TT_prod_6	0,000
	TT_prod_7	0,000
	TT_floor_9	0,000
	TT_gasOut_1	0,000
	PT_floor_6	0,000
	TT_floor_10	0,000
	TT_floor_11	0,000
	PT_floor_7	0,000
	PT_floor_8	0,000
	PT_floor_10	0,000

Table B.3: Summary of correlated variables in each dataset

Dataset	Correlated Variables	Correlation	Removed Variable
Dataset A	SC_fan_11, PT_floor_11	0,99	PT_floor_11
	PT_gasOut_1, PT_floor_4	0,95	PT_floor_4
	TT_prodOut_12, TT_prod_12	0,93	TT_prod_12
	SC_fan_11, SC_fan_10	0,92	None
	PT_floor_5, PT_floor_7	0,91	PT_floor_7
	SC_fan_10, PT_floor_11	0,90	PT_floor_11
Dataset B	SC_fan_11, PT_floor_11	0,99	PT_floor_11
	PT_gasOut_1, PT_floor_4	0,96	PT_floor_4
	TT_prodOut_12 , TT_prod_12	0,92	TT_prod_12
	SC_fan_11, SC_fan_10	0,92	None
	SC_fan_10, PT_floor_11	0,91	PT_floor_11
	TT_prodOut_12, TT_prod_11	0,89	TT_prod_11
	LIT_prod_5, LIT_prod_4	0,85	None
Dataset C	SC_fan_11, PT_floor_11	0,98	PT_floor_11
	PT_gasOut_1, PT_floor_4	0,98	PT_floor_4
	SC_fan_11, SC_fan_10	0,96	None
	TT_prod_8, TT_prod_5	0,95	TT_prod_8
	SC_fan_10, PT_floor_11	0,94	PT_floor_11
	SC_fan_10, PT_floor_10	0,92	PT_floor_10
	TT_prod_7, TT_prod_5	0,90	TT_prod_7
	PT_floor_11, PT_floor_10	0,88	PT_floor_10
	SC_fan_11, PT_floor_10	0,88	PT_floor_10
	TT_prod_6, TT_prod_5	0,89	TT_prod_5
	TT_prod_8, TT_prod_7	0,89	Both have been removed
	TT_gasOut_1, TT_prod_1	0,87	TT_prod_1
	PT_floor_7, PT_floor_5	0,87	PT_floor_5

Table B.4: Summary of correlated variables in the sliding window Dataset

Dataset	Correlated Variables	Correlation	Removed Variable
Sliding Window	SC_valv_4, SC_valv_5	0,94	None
	SC_valv_4, SC_valv_6	0,90	None
	SC_valv_12, SC_valv_5	0,91	None
	SC_valv_12, SC_valv_7	0,90	None
	PT_floor_10, SC_fan_10	0,95	PT_floor_10
	PT_floor_11, SC_fan_11	0,92	PT_floor_11
	PT_floor_4, PT_gasOut_1	1	PT_floor_4
	PT_floor_5, PT_floor_7	0,96	PT_floor_5
	PT_floor_6, PT_floor_7	0,99	PT_floor_6
	TT_prod_11, TT_prodOut_12	0,92	TT_prod_11
	TT_prod_12, TT_prodOut_12	0,99	TT_prod_12
	TT_floor_11, TT_prodOut_12	0,93	TT_floor_11

Table B.5: Minimum and maximum values for each variable in each group

Variables	**< 500**		**[500, 700[**		**[700, 900[**		**[900, 1100[**		**> 1100**	
	Min	Max	Min	Max	Min	Max	Min	Max	Min	Max
SC_fan_9	36	38	32	38	34	38	35	38	25	28
SC_fan_10	36	38	32	38	30	38	35	38	25	28
SC_fan_11	34	38	32	38	30	38	35	38	25	28
SC_valv_4	16	25	16	28	15	40	20	28	0	50
SC_valv_5	15	25	15	50	15	50	18	32	0	50
SC_valv_6	16	18	16	20	16	40	16	23	0	28
SC_valv_7	0	4	0	4	0	4	0	5	-8	6
SC_valv_8	11	15	11	15	11	15	11	15	0	19
SC_valv_12	7	15	6	18	8	17	14	18	-2	21
LIT_prod_4	37	48	37	52	3	49	37	52	2	51
LIT_prod_5	36	48	35	63	2	54	36	53	3	67
LIT_prod_6	36	53	34	54	22	59	38	54	1	44
LIT_prod_7	26	64	25	58	27	59	35	59	-1	55
LIT_prod_8	29	57	24	57	25	57	31	56	0	53
LIT_prod_12	19	41	15	38	20	40	21	38	9	40
TT_gasOut_1	73	75	73	78	73	99	73	76	73	100
PT_gasOut_1	-92	-18	-72	-17	-70	-12	-59	-31	-68	-22
IT_motor	352	374	330	393	268	392	311	374	186	357
TT_prod_1	74	80	73	80	72	99	73	81	75	100
TT_prod_4	63	101	62	104	62	103	63	104	62	104
TT_prod_5	105	107	105	106	105	106	105	106	105	111
TT_prod_6	104	107	102	107	102	107	103	107	102	116
TT_prod_7	106	107	106	107	105	107	105	108	105	113
TT_prod_8	106	107	106	107	106	107	105	108	102	109
TT_prod_9	49	63	45	69	45	61	45	65	44	75
TT_prod_10	39	55	45	57	47	55	45	59	41	57
TT_prod_11	38	55	35	59	38	55	35	57	33	55
TT_prod_12	27	44	28	52	29	47	27	40	24	44
TT_floor_9	99	129	93	129	103	129	94	123	60	127
TT_floor_10	25	127	31	126	81	127	29	127	25	126
TT_floor_11	20	109	21	110	21	109	21	109	21	105
PT_floor_4	-87	-11	-74	-10	-72	-13	-62	-26	-70	-15
PT_floor_5	19	110	-6	133	-34	122	11	138	-18	163
PT_floor_6	-167	600	-526	1292	-526	765	82	385	-20	398
PT_floor_7	380	662	302	832	130	752	295	295	36	798
PT_floor_8	53	53	53	53	53	53	53	53	-4	53
PT_floor_9	-14	13	-19	11	-12	11	-6	22	-31	10
PT_floor_10	307	368	3	370	243	374	309	371	162	372
PT_floor_11	380	353	259	357	230	356	302	357	146	360
PT_floor_12	339	444	213	450	256	446	347	432	222	432
TT_prodOut_12	26	40	26	45	28	41	28	40	23	41
TT_pump	55	65	56	65	57	62	56	61	41	58